VISIT…

ENCYCLOPEDIA OF

Multimedia Technology and Networking

Margherita Pagani

Encyclopedia of Multimedia Technology and Networking

Margherita Pagani
I-LAB Centre for Research on the Digital Economy, Bocconi University, Italy

IDEA GROUP REFERENCE
Hershey · London · Melbourne · Singapore

Acquisitions Editor: Renée Davies
Development Editor: Kristin Roth
Senior Managing Editor: Amanda Appicello
Managing Editor: Jennifer Neidig
Copy Editors: Julie LeBlanc, Shanelle Ramelb, Sue VanderHook and Jennifer Young
Typesetters: Diane Huskinson, Sara Reed and Larissa Zearfoss
Support Staff: Michelle Potter
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Idea Group Reference (an imprint of Idea Group Inc.)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@idea-group.com
Web site: http://www.idea-group-ref.com

and in the United Kingdom by
Idea Group Reference (an imprint of Idea Group Inc.)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 3313
Web site: http://www.eurospan.co.uk

Copyright © 2005 by Idea Group Inc. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher.

Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI of the trademark or registered trademark.

Library of Congress Cataloging-in-Publication Data

Encyclopedia of multimedia technology and networking / Margherita Pagani, ed.
p. cm.
Summary: "This encyclopedia offers a comprehensive knowledge of multimedia information technology from an economic and technological perspective"--Provided by publisher.
Includes bibliographical references and index.
ISBN 1-59140-561-0 (hard cover) -- ISBN 1-59140-796-6 (ebook)
1. Multimedia communications--Encyclopedias. I. Pagani, Margherita, 1971-
TK5105.15.E46 2005

2005005141

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this encyclopedia set is new, previously-unpublished material. The views expressed in this encyclopedia set are those of the authors, but not necessarily of the publisher.

Editorial Advisory Board

Raymond A. Hackney, *Manchester Metropolitan University, UK*

Leslie Leong, *Central Connecticut State University, USA*

Nadia Magnenat-Thalmann, *University of Geneva, Switzerland*

Lorenzo Peccati, *Bocconi University, Italy*

Steven John Simon, *Stetson School of Business and Economics - Mercer University, USA*

Andrew Targowski, *Western Michigan University, USA*

Nobuyoshi Terashima, *Waseda University, Japan*

Enrico Valdani, *Bocconi University, Italy*

List of Contributors

Ahmed, Ansary / *Open University of Malaysia, Malaysia*
Ajiferuke, Isola / *University of Western Ontario, Canada*
Akhtar, Shakil / *United Arab Emirates University, UAE*
Ally, Mohamed / *Athabasca University, Canada*
Angehrn, Albert A. / *Center for Advanced Learning Technologies, INSEAD, France*
Angelides, Marios C. / *Brunel University, UK*
Arbore, Alessandro / *Bocconi University, Italy*
Auld, Jonathan M. / *NovAtel Inc., Canada*
Baralou, Evangelia / *University of Sterling, Scotland*
Barolli, Leonard / *Fukuoka Institute of Technology, Japan*
Benrud, Erik / *American University, USA*
Bhattacharya, Sunand / *ITT Educational Services, Inc., USA*
Bignall, Robert J. / *Monash University, Australia*
Bradley, Randy V. / *Troy University, USA*
Buche, Mari W. / *Michigan Technological University, USA*
Butcher-Powell, Loreen Marie / *Bloomsburg University of Pennsylvania, USA*
Cannell, Jeremy C. / *Gannon University, USA*
Cardoso, Rui C. / *Universidade de Beira Interior, Portugal*
Cavallaro, Andrea / *Queen Mary, University of London, UK*
Chakrobartty, Shuvro / *Minnesota State University, USA*
Chan, Tom S. / *Southern New Hampshire University, USA*
Chbeir, Richard / *University of Bourgogne, France*
Chen, Jeanne / *HungKuang University, Taiwan*
Chen, Kuanchin / *Western Michigan University, USA*
Chen, Tung-Shou / *National Taichung Institute of Technology, Taiwan*
Cheng, Meng-Wen / *National Taichung Institute of Technology, Taiwan*
Chochcliouros, Ioannis P. / *Hellenic Telecommunications Organization S.A. (OTE), Greece*
Cirrincione, Armando / *SDA Bocconi School of Management, Italy*
Connaughton, Stacey L. / *Purdue University, USA*
Cragg, Paul B. / *University of Canterbury, New Zealand*
Cruz, Christophe / *University of Bourgogne, France*
da Silva, Elaine Quintino / *University of São Paulo, Brazil*
da Silva, Henrique J. A. / *Universidade de Coimbra, Portugal*
Danenberg, James O. / *Western Michigan University, USA*
de Abreu Moreira, Dilvan / *University of São Paulo, Brazil*
de Amescua, Antonio / *Carlos III Technical University of Madrid, Spain*
Dellas, Fabien / *University of Geneva, Switzerland*
Dhar, Subhankar / *San Jose State University, USA*
Di Giacomo, Thomas / *University of Geneva, Switzerland*

Diaz, Ing. Carlos / *University of Alcala, Spain*
Dunning, Jeremy / *Indiana University, USA*
Duthler, Kirk W. / *The University of North Carolina at Charlotte, USA*
El-Gayar, Omar / *Dakota State University, USA*
Esmahi, Larbi / *Athabasca University, Canada*
Esteban, Luis A. / *Carlos III Technical University of Madrid, Spain*
Falk, Louis K. / *Youngstown State University, USA*
Farag, Waleed E. / *Zagazig University, Egypt*
Fleming, Stewart T. / *University of Otago, New Zealand*
Fraunholz, Bardo / *Deakin University, Australia*
Freeman, Ina / *University of Birmingham, UK*
Freire, Mário M. / *Universidade de Beira Interior, Portugal*
Galanxhi-Janaqi, Holtjona / *University of Nebraska-Lincoln, USA*
Gao, Yuan / *Ramapo College of New Jersey, USA*
García, Luis / *Carlos III Technical University of Madrid, Spain*
Ghanem, Hassan / *Verizon, USA*
Gibbert, Michael / *Bocconi University, Italy*
Gilbert, A. Lee / *Nanyang Business School, Singapore*
Goh, Tiong-Thye / *Victoria University of Wellington, New Zealand*
Grahn, Kaj J. / *Arcada Polytechnic, Finland*
Grover, Akshay / *Brigham Young University, USA*
Guan, Sheng-Uei / *National University of Singapore, Singapore*
Gupta, P. / *Indian Institute of Technology Kanpur, India*
Gurău, Călin / *Centre d'Études et de Recherche sur les Organisations et la Management (CEROM), France*
Gutiérrez, Jairo A. / *The University of Auckland, New Zealand*
Hackbarth, Klaus D. / *University of Cantabria, Spain*
Hagenhoff, Svenja / *Georg-August-University of Goettingen, Germany*
Handzic, Meliha / *Sarajevo School of Science and Technology, BiH, Croatia*
Hentea, Mariana / *Southwestern Oklahoma State University, USA*
Heywood, Malcolm I. / *Dalhousie University, Canada*
Hin, Leo Tan Wee / *Singapore National Academy of Science and Nanyang Technological University, Singapore*
Hosszú, Gábor / *Budapest University of Technology and Economics, Turkey*
Hu, Wen-Chen / *University of North Dakota, USA*
Hughes, Jerald / *Baruch College of the City University of New York, USA*
Hulicki, Zbigniew / *AGH University of Science and Technology, Poland*
Hurson, Ali R. / *The Pennsylvania State University, USA*
Iossifides, Athanassios C. / *COSMOTE S.A., Greece*
Ishaya, Tanko / *The University of Hull, UK*
Janczewski, Lech J. / *The University of Auckland, New Zealand*
Joslin, Chris / *University of Geneva, Switzerland*
Jovanovic-Dolecek, Gordana / *INAOE, Mexico*
Jung, Jürgen / *Uni Duisburg-Essen, Germany*
Kacimi, Mouna / *University of Bourgogne, France*
Kanellis, Panagiotis / *National and Kapodistrian University of Athens, Greece*
Karaboulas, Dimitrios / *University of Patras, Greece*
Karlsson, Jonny / *Arcada Polytechnic, Finland*
Karoui, Kamel / *Institut National des Sciences Appliquées de Tunis, Tunisia*

Kaspar, Christian / *Georg-August-University of Goettingen, Germany*
Kaur, Abtar / *Open University of Malaysia, Malaysia*
Kaushik, A.K. / *Electronic Niketan, India*
Kayacik, H. Gunes / *Dalhousie University, Canada*
Kelic, Andjelka / *Massachusetts Institute of Technology, USA*
Kemper Littman, Marlyn / *Nova Southeastern University, USA*
Kinshuk / *Massey University, New Zealand*
Knight, Linda V. / *DePaul University, USA*
Kontolemakis, George / *National and Kapodistrian University of Athens, Greece*
Kotsopoulos, Stavros A. / *University of Patras, Greece*
Kou, Weidong / *Xidian University, PR China*
Koumaras, Harilaos / *University of Athens, Greece*
Kourtis, Anastasios / *Institute of Informatics and Telecommunications NCSR Demokritos, Greece*
Koyama, Akio / *Yamagata University, Japan*
Kwok, Percy Lai-yin / *Chinese University of Hong Kong, China*
Labruyere, Jean-Philippe P. / *DePaul University, USA*
Lalopoulos, George K. / *Hellenic Telecommunications Organization S.A. (OTE), Greece*
Lang, Karl Reiner / *Baruch College of the City University of New York, USA*
Lang, Michael / *National University of Ireland, Galway, Ireland*
Larkin, Jeff / *Brigham Young University, USA*
Lawson-Body, Assion / *University of North Dakota, USA*
Lee, Chung-wei / *Auburn University, USA*
Lee, Maria Ruey-Yuan / *Shih-Chien University, Taiwan*
Li, Chang-Tsun / *University of Warwick, UK*
Li, Qing / *City University of Hong Kong, China*
Liehr, Marcus / *University of Hohenheim, Germany*
Lin, Joanne Chia Yi / *The University of New South Wales, Australia*
Lorenz, Pascal / *University of Haute Alsace, France*
Louvros, Spiros / *COSMOTE S.A., Greece*
Lowry, Paul Benjamin / *Brigham Young University, USA*
Lumsden, Joanna / *National Research Council of Canada IIT e-Business, Canada*
Luo, Xin / *Mississippi State University, USA*
Ma, Keh-Jian / *National Taichung Institute of Technology, Taiwan*
Madsen, Chris / *Brigham Young University, USA*
Maggioni, Mario A. / *Università Cattolica di Milano, Italy*
Magnenat-Thalmann, Nadia / *University of Geneva, Switzerland*
Magni, Massimo / *Bocconi University, Italy*
Maris, Jo-Mae B. / *Northern Arizona University, USA*
Markus, Alexander / *University of Western Ontario, Canada*
Martakos, Drakoulis / *National and Kapodistrian University of Athens, Greece*
Mbarika, Victor / *Southern University and A&M College, USA*
McManus, Patricia / *Edith Cowan University, Australia*
Melliar-Smith, P. M. / *University of California, Santa Barbara, USA*
Mills, Annette M. / *University of Canterbury, New Zealand*
Mitchell, Mark / *Brigham Young University, USA*
Mohamedally, Dean / *City University London, UK*
Monteiro, Paulo P. / *SIEMENS S.A. and Universidade de Aveiro, Portugal*
Morabito, Vincenzo / *Bocconi University, Italy*
Moser, L. E. / *University of California, Santa Barbara, USA*

Moyes, Aaron / *Brigham Young University, USA*
Mundy, Darren P. / *University of Hull, UK*
Murphy, Peter / *Victoria University of Wellington, New Zealand*
Nah, Fiona Fui-Hoon / *University of Nebraska-Lincoln, USA*
Nandavadekar, Vilas D. / *University of Pune, India*
Neveu, Marc / *University of Burgundy, France*
Ngoh, Lek Heng / *Institute for Infocomm Research, A*STAR, Singapore*
Nicolle, Christophe / *University of Bourgogne, France*
Nur, Mohammad M. / *Minnesota State University, USA*
Nur Zincir-Heywood, A. / *Dalhousie University, Canada*
O'Dea, Michael / *University of Hull, UK*
O'Hagan, Minako / *Dublin City University, Ireland*
Olla, Phillip / *Brunel University, UK*
Otenko, Oleksandr / *University of Kent, UK*
Pace, Stefano / *Bocconi University, Italy*
Pagani, Margherita / *Bocconi University, Italy*
Pai, Feng Yu / *Shih-Chien University, Taiwan*
Panjala, Shashidhar / *Gannon University, USA*
Pantic, Maja / *Delft University of Technology, The Netherlands*
Pereira, Rui G. / *Universidade de Beira Interior, Portugal*
Petrie, Helen / *City University London, UK*
Poole, Marshall Scott / *Texas A&M University, USA*
Portilla, J. Antonio / *University of Alcala, Spain*
Portougal, Victor / *The University of Auckland, New Zealand*
Prata, Alcina / *Higher School of Management Sciences, Portugal*
Proserpio, Luigi / *Bocconi University, Italy*
Provera, Bernardino / *Bocconi University, Italy*
Pulkkis, Göran / *Arcada Polytechnic, Finland*
Rahman, Hakikur / *SDNP, Bangladesh*
Raisinghani, Mahesh S. / *Texas Woman's University, USA*
Rajasingham, Lalita / *Victoria University of Wellington, New Zealand*
Raju, P.K. / *Auburn University, USA*
Ratnasingam, Pauline / *Central Missouri State University, USA*
Ripamonti, Laura Anna / *Università degli Studi di Milano, Italy*
Robins, Wiliam / *Brigham Young University, USA*
Rodrigues, Joel J. P. C. / *Universidade da Beira Interior, Portugal*
Rotvold, Glenda / *University of North Dakota, USA*
Rotvold, Justin / *Techwise Solutions, LLC, USA*
Rowe, Neil C. / *U.S. Naval Postgraduate School, USA*
Roy, Abhijit / *Loyola College in Maryland, USA*
Ruela, Jose / *Faculdade de Engennaria da Universidade do Porto (FEUP), Portugal*
Sánchez-Segura, Maria-Isabel / *Carlos III Technical University of Madrid, Spain*
Sankar, Chetan S. / *Auburn University, USA*
Schizas, Christos / *University of Cyprus, Cyprus*
Shankar P., Jaya / *Institute for Infocomm Research, A*STAR, Singapore*
Shepherd, Jill / *Simon Fraser University, Canada*
Shuaib, Khaled A. / *United Arab Emirates University, UAE*
Singh, Richa / *Indian Institute of Technology Kanpur, India*
Singh, Shawren / *University of South Africa, South Africa*

Sockel, Hy / *Youngstown State University, USA*
Sofokleous, Anastasis / *Brunel University, UK*
Sonwalkar, Nishikant / *Massachusetts Institute of Technology, USA*
Spiliopoulou-Chochliourou, Anastasia S. / *Hellenic Telecommunications Organization S.A. (OTE), Greece*
St.Amant, Kirk / *Texas Tech University, USA*
Standing, Craig / *Edith Cowan University, Australia*
Stephens, Jackson / *Brigham Young University, USA*
Stern, Tziporah / *Baruch College, CUNY, USA*
Still, Brian / *Texas Tech University, USA*
Subramaniam, R. / *Singapore National Academy of Science and Nanyang Technological University, Singapore*
Sun, Jun / *Texas A&M University, USA*
Suraweera, Theekshana / *University of Canterbury, New Zealand*
Swierzowicz, Janusz / *Rzeszow University of Technology, Poland*
Syed, Mahbubur R. / *Minnesota State University, USA*
Szabados, Anna / *Mission College, USA*
Tan, Christopher Yew-Gee / *University of South Australia, Australia*
Tandekar, Kanchana / *Dakota State University, USA*
Tassabehji, Rana / *University of Bradford, UK*
Terashima, Nobuyoshi / *Waseda University, Japan*
Tiffin, John / *Victoria University of Wellington, New Zealand*
Ting, Wayne / *The University of Auckland, New Zealand*
Todorova, Nelly / *University of Canterbury, New Zealand*
Tong, Carrison KS / *Pamela Youde Nethersole Eastern Hospital and Tseung Kwan O Hospital, Hong Kong*
Torrisi-Steele, Geraldine / *Griffith University, Australia*
Uberti, Teodora Erika / *Università Cattolica di Milano, Italy*
Unnithan, Chandana / *Deakin University, Australia*
Vatsa, Mayank / *Indian Institute of Technology Kanpur, India*
Vician, Chelley / *Michigan Technological University, USA*
Vitolo, Theresa M. / *Gannon University, USA*
Voeth, Markus / *University of Hohenheim, Germany*
Volino, Pascal / *University of Geneva, Switzerland*
Wang, Pin-Hsin / *National Taichung Institute of Technology, Taiwan*
Warkentin, Merrill / *Mississippi State University, USA*
Wei, Chia-Hung / *University of Warwick, UK*
Wilson, Sean / *Brigham Young University, USA*
Wong, Eric TT / *The Hong Kong Polytechnic University, Hong Kong*
Wong-MingJi, Diana J. / *Eastern Michigan University, USA*
Wright, Carol / *Pennsylvania State University, USA*
Yang, Bo / *The Pennsylvania State University, USA*
Yang, Jun / *Carnegie Mellon University, USA*
Yetongnon, Kokou / *University of Bourgogne, France*
Yusof, Shafiz A. Mohd / *Syracuse University, USA*
Zakaria, Norhayati / *Syracuse University, USA*
Zaphiris, Panayiotis / *City University London, UK*
Zhuang, Yueting / *Zhejiang University, China*
Zwitserloot, Reinier / *Delft University of Technology, The Netherlands*

Contents

by Volume

VOLUME I

Adoption of Communication Products and the Individual Critical Mass / *Markus Voeth and Marcus Liehr* 1

Affective Computing / *Maja Pantic* 8

Agent Frameworks / *Reinier Zwitserloot and Maja Pantic* 15

Application of Genetic Algorithms for QoS Routing in Broadband Networks / *Leonard Barolli and Akio Koyama* 22

Application Service Providers / *Vincenzo Morabito and Bernardino Provera* 31

Assessing Digital Video Data Similarity / *Waleed E. Farag* 36

Asymmetric Digital Subscriber Line / *Leo Tan Wee Hin and R. Subramaniam* 42

ATM Technology and E-Learning Initiatives / *Marlyn Kemper Littman* 49

Biometric Technologies / *Mayank Vatsa, Richa Singh, P. Gupta and A.K. Kaushik* 56

Biometrics Security / *Stewart T. Fleming* 63

Biometrics, A Critical Consideration in Information Security Management / *Paul Benjamin Lowry, Jackson Stephens, Aaron Moyes, Sean Wilson and Mark Mitchell* 69

Broadband Solutions for Residential Customers / *Mariana Hentea* 76

Challenges and Perspectives for Web-Based Applications in Organizations / *George K. Lalopoulos, Ioannis P. Chochliouros and Anastasia S. Spiliopoulou-Chochliourou* 82

Collaborative Web-Based Learning Community / *Percy Kwok Lai-yin and Christopher Tan Yew-Gee* 89

Constructing a Globalized E-Commerce Site / *Tom S. Chan* 96

Consumer Attitude in Electronic Commerce / *Yuan Gao* 102

Content Repurposing for Small Devices / *Neil C. Rowe* 110

Content-Based Multimedia Retrieval / *Chia-Hung Wei and Chang-Tsun Li* 116

Context-Awareness in Mobile Commerce / *Jun Sun and Marshall Scott Poole* 123

Core Principles of Educational Multimedia / *Geraldine Torrisi-Steele* 130

Corporate Conferencing / *Vilas D. Nandavadekar* 137

Cost Models for Telecommunication Networks and Their Application to GSM Systems / *Klaus D. Hackbarth, J. Antonio Portilla and Ing. Carlos Diaz* 143

Critical Issues in Global Navigation Satellite Systems / *Ina Freeman and Jonathan M. Auld* 151

Dark Optical Fibre as a Modern Solution for Broadband Networked Cities / *Ioannis P. Chochliouros, Anastasia S. Spiliopoulou-Chochliourou and George K. Lalopoulos* 158

Decision Making Process of Integrating Wireless Technology into Organizations, The / *Assion Lawson-Body, Glenda Rotvold and Justin Rotvold* 165

Designing Web-Based Hypermedia Systems / *Michael Lang* 173

Digital Filters / *Gordana Jovanovic-Dolecek* 180

Digital Video Broadcasting (DVB) Applications / *Ioannis P. Chochliouros, Anastasia S. Spiliopoulou-Chochliourou and George K. Lalopoulos* 197

Digital Watermarking Based on Neural Network Technology for Grayscale Images / *Jeanne Chen, Tung-Shou Chen, Keh-Jian Ma and Pin-Hsin Wang* 204

Digital Watermarking for Multimedia Security Management / *Chang-Tsun Li* 213

Distance Education Delivery / *Carol Wright* 219

Distanced Leadership and Multimedia / *Stacey L. Connaughton* 226

Dynamics of Virtual Teams, The / *Norhayati Zakaria and Shafiz A. Mohd Yusof* 233

E-Commerce and Usability / *Shawren Singh* 242

Educational Technology Standards / *Michael O'Dea* 247

Efficient Method for Image Indexing in Medical Application / *Richard Chbeir* 257

Elaboration Likelihood Model and Web-Based Persuasion, The / *Kirk W. Duthler* 265

E-Learning and Multimedia Databases / *Theresa M. Vitolo, Shashidhar Panjala and Jeremy C. Cannell* 271

Electronic Commerce Technologies Management / *Shawren Singh* 278

Ethernet Passive Optical Networks / *Mário M. Freire, Paulo P. Monteiro, Henrique J. A. da Silva and Jose Ruela* 283

Evolution of GSM Network Technology / *Phillip Olla* 290

Evolution of Mobile Commerce Applications / *George K. Lalopoulos, Ioannis P. Chochcliouros and Anastasia S. Spiliopoulou-Chochliourou* 295

Exploiting Captions for Multimedia Data Mining / *Neil C. Rowe* 302

Face for Interface / *Maja Pantic* 308

FDD Techniques Towards the Multimedia Era / *Athanassios C. Iossifides, Spiros Louvros and Stavros A. Kotsopoulos* 315

Fiber to the Premises / *Mahesh S. Raisinghani and Hassan Ghanem* 324

Fiber-to-the-Home Technologies and Standards / *Andjelka Kelic* 329

From Communities to Mobile Communities of Values / *Patricia McManus and Craig Standing* 336

Future of M-Interaction, The / *Joanna Lumsden* 342

Global Navigation Satellite Systems / *Phillip Olla* 348

Going Virtual / *Evangelia Baralou and Jill Stepherd* 353

Heterogeneous Wireless Networks Using a Wireless ATM Platform / *Spiros Louvros, Dimitrios Karaboulas, Athanassios C. Iossifides and Stavros A. Kotsopoulos* 359

HyperReality / *Nobuyoshi Terashima* 368

Improving Student Interaction with Internet and Peer Review / *Dilvan de Abreu Moreira and Elaine Quintino da Silva* 375

Information Hiding, Digital Watermarking and Steganography / *Kuanchin Chen* 382

Information Security Management / *Mariana Hentea* 390

Information Security Management in Picture-Archiving and Communication Systems for the Healthcare Industry / *Carrison KS Tong and Eric TT Wong* 396

Information Security Threats / *Rana Tassabehji* 404

Information Systems Strategic Alignment in Small Firms / *Paul B. Cragg and Nelly Todorova* 411

Information Technology and Virtual Communities / *Chelley Vician and Mari W. Buche* 417

Integrated Platform for Networked and User-Oriented Virtual Clothing / *Pascal Volino, Thomas Di Giacomo, Fabien Dellas and Nadia Magnenat-Thalmann* 424

Interactive Digital Television / *Margherita Pagani* 428

Interactive Memex / *Sheng-Uei Guan* 437

Interactive Multimedia Technologies for Distance Education in Developing Countries / *Hakikur Rahman* 447

Interactive Multimedia Technologies for Distance Education Systems / *Hakikur Rahman* 454

International Virtual Offices / *Kirk St.Amant* 461

Internet Adoption by Small Firms / *Paul B. Cragg and Annette M. Mills* 467

Internet Privacy from the Individual and Business Perspectives / *Tziporah Stern* 475

Internet Privacy Issues / *Hy Sockel and Kuanchin Chen* 480

Interoperable Learning Objects Management / *Tanko Ishaya* 486

Intrusion Detection Systems / *H. Gunes Kayacik, A. Nur Zincir-Heywood and Malcolm I. Heywood* 494

Investment Strategy for Integrating Wireless Technology into Organizations / *Assion Lawson-Body* 500

IT Management Practices in Small Firms / *Paul B. Cragg and Theekshana Suraweera* 507

iTV Guidelines / *Alcina Prata* 512

Leadership Competencies for Managing Global Virtual Teams / *Diana J. Wong-MingJi* 519

Learning Networks / *Albert A. Angehrn and Michael Gibbert* 526

Learning through Business Games / *Luigi Proserpio and Massimo Magni* 532

Local Loop Unbundling / *Alessandro Arbore* 538

Local Loop Unbundling Measures and Policies in the European Union / *Ioannis P. Chochliouros, Anastasia S. Spiliopoulou-Chochliourou and George K. Lalopoulos* 547

VOLUME II

Making Money with Open-Source Business Initiatives / *Paul Benjamin Lowry, Akshay Grover, Chris Madsen, Jeff Larkin and William Robins* 555

Malware and Antivirus Procedures / *Xin Luo and Merrill Warkentin* 562

Measuring the Potential for IT Convergence at Macro Level / *Margherita Pagani* 571

Message-Based Service in Taiwan / *Maria Ruey-Yuan Lee and Feng Yu Pai* 579

Methods of Research in Virtual Communities / *Stefano Pace* 585

Migration to IP Telephony / *Khaled A. Shuaib* 593

Mobile Ad Hoc Network / *Subhankar Dhar* 601

Mobile Agents / *Kamel Karoui* 608

Mobile Commerce Security and Payment / *Chung-wei Lee, Weidong Kou and Wen-Chen Hu* 615

Mobile Computing for M-Commerce / *Anastasis Sofokleous, Marios C. Angelides and Christos Schizas* 622

Mobile Location Based Services / *Bardo Fraunholz, Jürgen Jung and Chandana Unnithan* 629

Mobile Multimedia for Commerce / *P. M. Melliar-Smith and L. E. Moser* 638

Mobile Radio Technologies / *Christian Kaspar and Svenja Hagenhoff* 645

Mobility over Heterogeneous Wireless Networks / *Lek Heng Ngoh and Jaya Shankar P.* 652

Modeling Interactive Distributed Multimedia Applications / *Sheng-Uei Guan* 660

Modelling eCRM Systems with the Unified Modelling Language / *Călin Gurău* 667

Multimedia Communication Services on Digital TV Platforms / *Zbigniew Hulicki* 678

Multimedia Content Representation Technologies / *Ali R. Hurson and Bo Yang* 687

Multimedia Data Mining Concept / *Janusz Swierzowicz* 696

Multimedia Information Design for Mobile Devices / *Mohamed Ally* 704

Multimedia Information Retrieval at a Crossroad / *Qing Li, Jun Yang, and Yueting Zhuang* 710

Multimedia Instructional Materials in MIS Classrooms / *Randy V. Bradley, Victor Mbarika, Chetan S. Sankar and P.K. Raju* 717

Multimedia Interactivity on the Internet / *Omar El-Gayar, Kuanchin Chen and Kanchana Tandekar* 724

Multimedia Proxy Cache Architectures / *Mouna Kacimi, Richard Chbeir and Kokou Yetongnon* 731

Multimedia Technologies in Education / *Armando Cirrincione* 737

N-Dimensional Geometry and Kinaesthetic Space of the Internet, The / *Peter Murphy* 742

Network Intrusion Tracking for DoS Attacks / *Mahbubur R. Syed, Mohammad M. Nur and Robert J. Bignall* 748

Network-Based Information System Model for Research / *Jo-Mae B. Maris* 756

New Block Data Hiding Method for the Binary Image, A / *Jeanne Chen, Tung-Shou Chen and Meng-Wen Cheng* 762

Objective Measurement of Perceived QoS for Homogeneous MPEG-4 Video Content / *Harilaos Koumaras, Drakoulis Martakos and Anastasios Kourtis* 770

Online Discussion and Student Success in Web-Based Education, The / *Erik Benrud* 778

Open Source Intellectual Property Rights / *Stewart T. Fleming* 785

Open Source Software and International Outsourcing / *Kirk St.Amant and Brian Still* 791

Optical Burst Switching / *Joel J. P. C. Rodrigues, Mário M. Freire, Paulo P. Monteiro and Pascal Lorenz* 799

Peer-to-Peer Filesharing Systems for Digital Media / *Jerald Hughes and Karl Reiner Lang* 807

Personalized Web-Based Learning Services / *Larbi Esmahi* 814

Picture Archiving and Communication System in Health Care / *Carrison KS Tong and Eric TT Wong* 821

Plastic Optical Fiber Applications / *Spiros Louvros, Athanassios C. Iossifides, Dimitrios Karaboulas and Stavros A. Kotsopoulos* 829

Potentials of Information Technology in Building Virtual Communities / *Isola Ajiferuke and Alexander Markus* 836

Principles for Managing Information Security / *Rana Tassabehji* 842

Privilege Management Infrastructure / *Darren P. Mundy and Oleksandr Otenko* 849

Production, Delivery and Playback of 3D Graphics / *Thomas Di Giacomo, Chris Joslin, and Nadia Magnenat-Thalmann* 855

Public Opinion and the Internet / *Peter Murphy* 863

Quality of Service Issues Associated with Internet Protocols / *Jairo A. Gutiérrez and Wayne Ting* 869

Reliability Issues of the Multicast-Based Mediacommunication / *Gábor Hosszú* 875

Re-Purposeable Learning Objects Based on Teaching and Learning Styles / *Abtar Kaur, Jeremy Dunning, Sunand Bhattacharya and Ansary Ahmed* 882

Risk-Control Framework for E-Marketplace Participation, A / *Pauline Ratnasingam* 887

Road Map to Information Security Management / *Lech J. Janczewski and Victor Portougal* 895

Security Laboratory Design and Implementation / *Linda V. Knight and Jean-Philippe P. Labruyere* 903

Security Vulnerabilities and Exposures in Internet Systems and Services / *Rui C. Cardoso and Mário M. Freire* 910

Semantic Web / *Rui G. Pereira and Mário M. Freire* 917

Software Ad Hoc for E-Learning / *Maria-Isabel Sánchez-Segura, Antonio de Amescua, Luis García and Luis A. Esteban* 925

Supporting Online Communities with Technological Infrastructures / *Laura Anna Ripamonti* 937

Teletranslation / *Minako O'Hagan* 945

Telework Information Security / *Loreen Marie Butcher-Powell* 951

Text-to-Speech Synthesis / *Mahbubur R. Syed, Shuvro Chakrobartty and Robert J. Bignall* 957

2G-4G Networks / *Shakil Akhtar* 964

Type Justified / *Anna Szabados and Nishikant Sonwalkar* 974

Ubiquitous Commerce / *Holtjona Galanxhi-Janaqi and Fiona Fui-Hoon Nah* 980

Understanding the Out-of-the-Box Experience / *A. Lee Gilbert* 985

Unified Information Security Management Plan, A / *Mari W. Buche and Chelley Vician* 993

Universal Multimedia Access / *Andrea Cavallaro* 1001

Usability / *Shawren Singh* 1008

Usability Assessment in Mobile Computing and Commerce / *Kuanchin Chen, Hy Sockel and Louis K. Falk* 1014

User-Centered Mobile Computing / *Dean Mohamedally, Panayiotis Zaphiris and Helen Petrie* 1021

Using Semantics to Manage 3D Scenes in Web Platforms / *Christophe Cruz, Christophe Nicolle and Marc Neveu* 1027

Virtual Communities / *George Kontolemakis, Panagiotis Kanellis and Drakoulis Martakos* 1033

Virtual Communities on the Internet / *Abhijit Roy* 1040

Virtual Knowledge Space and Learning / *Meliha Handzic and Joanne Chia Yi Lin* 1047

Virtual Learning Communities / *Stewart T. Fleming* 1055

Virtual Reality and HyperReality Technologies in Universities / *Lalita Rajasingham and John Tiffin* 1064

Web Content Adaptation Frameworks and Techniques / *Tiong-Thye Goh and Kinshuk* 1070

Web Site Usability / *Louis K. Falk and Hy Sockel* 1078

Web-Based Learning / *James O. Danenberg and Kuanchin Chen* 1084

Webmetrics / *Mario A. Maggioni and Teodora Erika Uberti* 1091

Wireless Emergency Services / *Jun Sun* 1096

WLAN Security Management / *Göran Pulkkis, Kaj J. Grahn, and Jonny Karlsson* 1104

Foreword

Multimedia technology and networking are changing at a remarkable rate. Despite the telecoms crash of 2001, innovation in networking applications, technologies, and services has continued unabated. The exponential growth of the Internet, the explosion of mobile communications, the rapid emergence of electronic commerce, the restructuring of businesses, and the contribution of digital industries to growth and employment, are just a few of the current features of the emerging digital economy.

The *Encyclopedia of Multimedia Technology and Networking* captures a vast array of the components and dimensions of this dynamic sector of the world economy. Professor Margherita Pagani and her editorial board have done a remarkable job at compiling such a rich collection of perspectives on this fast moving domain. The encyclopaedia's scope and content will provide scholars, researchers and professionals with much current information about concepts, issues, trends and technologies in this rapid evolving industrial sector.

Multimedia technologies and networking are at the heart of the current debate about economic growth and performance in advanced economies. The pervasive nature of the technological change and its widespread diffusion has profoundly altered the ways in which businesses and consumers interact. As IT continues to enter workplaces, homes and learning institutions, many aspects of work and leisure are changing radically. The rapid pace of technological change and the growing connectivity that IT makes possible have resulted in a wealth of new products, new markets and new business models. However, these changes also bring new risks, new challenges, and new concerns.

In the multimedia and technology networks area broadband-based communication and entertainment services are helping consumer and business users to conduct business more effectively, serve customers faster, and organise their time more effectively. In fact, multimedia technologies and networks have a strong impact on all economic activity. Exponential growth in processing power, falling information costs and network effects have allowed productivity gains, enhanced innovation, and stimulated further technical change in all sectors from the most technology intensive to the most traditional. Broadband communications and entertainment services are helping consumer and business users conduct their business more effectively, serve customers faster, organise their time more effectively, and enrich options for their leisure time.

At MIT, I serve as co-director of the Communications Futures Program, which spans the Sloan School of Management, the Engineering School, and the Media Lab at the Massachusetts Institute of Technology (USA). By examining technology dynamics, business dynamics, and policy dynamics in the communications industry, we seek to build capabilities for roadmapping the upcoming changes in the vast communications value chain. We also seek to develop next-generation technological and business innovations that can create more value in the industry.

Furthermore, we hope that gaining a deeper understanding of the dynamics in communications will help us not only to make useful contributions to that field, but also to understand better the general principles that drive industry and technology dynamics. Biologists study fruit flies because their fast rates of evolution permit rapid learning that can then be applied to understanding the genetics of slower clockspeed species, like humans. We think of the communications industry as the industrial equivalent of a fruit fly; that is, a fast

clockspeed industry whose dynamics may help us understand better the dynamic principles that drive many industries.

Convergence is among the core features of information society developments. This phenomenon needs to be analyzed from multiple dimensions: technological, economic, financial, regulatory, social, and political. The integrative approach adopted in this encyclopaedia to analyze multimedia and technology networking is particularly welcome and highly complementary to the approach embraced by our work at MIT.

I am pleased to be able to recommend this encyclopedia to readers, be they looking for substantive material on knowledge strategy, or looking to understand critical issues related to multimedia technology and networking.

Professor Charles H. Fine
Massachusetts Institute of Technology
Sloan School of Management
Cambridge, October 2004

Preface

The term encyclopedia comes from the Greek words εγκύκλιος παιδεία , enkyklios paideia ("in a circle of instruction").

The purpose of the *Encyclopedia of Multimedia Technology and Networking* is to offer a written compendium of human knowledge related to the emerging multimedia digital metamarket.

Multimedia technology, networks and online interactive multimedia services are taking advantage of a series of radical innovations in converging fields, such as the digitization of signals, satellite and fibre optic based transmission systems, algorithms for signal compression and control, switching and storage devices, and others, whose combination has a supra-additive synergistic effect.

The emergence of online interactive multimedia (OIM) services can be described as a new technological paradigm. They can be defined by a class of new techno economic problems, a new pool of technologies (techniques, competencies and rules), and a set of shared assumptions. The core of such a major shift in the evolution of information and communications services is the service provision function. This shirt occurs even if the supply of an online interactive multimedia service needs a wide collection of assets and capabilities pertaining also to information contents, network infrastructure, software, communication equipment and terminals.

By zooming in on the operators of telecommunications networks (common carriers or telecoms), it is shown that though leveraging a few peculiar capabilities in the technological and managerial spheres, they are trying to develop lacking assets and competencies through the set-up of a network of collaborative relations with firms in converging industries (mainly software producers, service providers, broadcasters, and media firms). This emerging digital marketplace is constantly expanding.

As new platforms and delivery mechanisms rapidly roll out, the value of content increases, presenting content owners with both risks and opportunities. In addition, rather than purely addressing the technical challenge of the Internet, wireless and interactive digital television, much more emphasis is now being given to commercial and marketing issues. Companies are much more focused on the creation of consistent and compelling user experiences.

The use of multimedia technologies as the core driving element in converging markets and virtual corporate structures will compel considerable economic and social change.

Set within the framework of IT as a strategic resource, many important changes have taken place over the last years that will force us to change the way multimedia networks develop services for their users.

- The change in the expectations of users, leading to new rapid development and implementation techniques;
- The launch of next generation networks and handsets;
- The rapid pace at which new technologies (software and hardware) are introduced;
- Modularization of hardware and software, emphasizing object assembly and processing (client server computing);
- Development of non-procedural languages (visual and object oriented programming);

- An imbalance between network operators and independent application developers in the value network for the provision of network dependent services;
- Telecommunications integrated into, and inseparable from, the computing environment;
- Need for integration of seemingly incompatible diverse technologies.

The force behind these realities is the strategic use of IT. Strategic management which takes into consideration the basic transformation processes of this sector will be a substantial success factor in securing a competitive advantage within this deciding future market. The change from an industrial to an information society connected therewith, will above all else be affected by the dynamics of technological developments.

This strategic perspective manifests itself in these work attributes:

- an appreciation of IT within the context of business value;
- a view of information as a critical resource to be managed and developed as an asset;
- a continuing search for opportunities to exploit information technology for competitive advantage;
- uncovering opportunities for process redesign;
- concern for aligning IT with organizational goals;
- a continuing re-evaluation of work assignments for added value;
- skill in adapting quickly to appropriate new technologies;
- an object/modular orientation for technical flexibility and speed in deployment.

Accelerating economic, technological, social, and environmental change challenges managers and policy makers to learn at increasing rates, while at the same time the complexity of the systems in which we live is growing.

Effective decision making and learning in a world of growing *dynamic complexity* requires us to develop tools to understand how the structure of complex systems creates their behaviour.

THE EMERGING MULTIMEDIA MARKET

The convergence of information and communication technology has lead to the development of a variety of new media platforms that offer a set of services to a community of participants. These platforms are defined as media which enable the exchange of information or other objects such as goods and services (Schmid, 1999).

Media can be defined as information and communication spaces, which based on innovative information and communication technology (ICT), supports content creation, management and exchange within a community of agents. Agents can be organizations, humans, or artificial agents (i.e., software agents).

The multimedia metamarket—generated by the progressive process of convergence involving the television, informatics and telecommunication industries—comes to represent the «strategic field of action» of this study.

According to this perspective, telecommunications, office equipment, consumer electronics, media, and computers were separate and distinct industries through the 1990s. They offered different services with different methods of delivery. But as the computer became an “information appliance”, businesses moved to take advantage of emerging digital technologies, virtual reality, and industry boundaries blurred.

As a result of the convergence process, we cannot, therefore, talk about separate and different industries and sectors (telecommunications, digital television, and informatics). Such sectors are propelled towards an actual merging of different technologies, supplied services and the users’ categories being reached. A great ICT metamarket is thus originated.

Multimedia finds its application in various areas including, but not limited to, education, entertainment, engineering, medicine, mathematics, and scientific research.

In education, multimedia is used to produce computer based training courses.

Multimedia is heavily used in the entertainment industry, especially to develop special effects in movies and animation for cartoon characters. Multimedia games such as software programs available either as CD-ROMs or online are a popular pastime.

In engineering, especially mechanical and automobile engineering, multimedia is primarily used for designing a machinery or automobile. This lets an engineer view a product from various perspectives, zoom critical parts and do other manipulations, before actually producing it. This is known as computer aided design (CAD).

In medicine, doctors can get trained by looking at a virtual surgery.

In mathematical and scientific research, multimedia is mainly used for modelling and simulation. For example, a scientist can look at a molecular model of a particular substance and manipulate it to arrive at a new substance.

Multimedia technologies and networking are at the heart of the current debate about economic growth and performance in advanced economies.

ORGANIZATION OF THIS ENCYCLOPEDIA

The goal of the *Encyclopedia of Multimedia Technology and Networking* is to improve our understanding of multimedia and digital technologies adopting an integrative approach.

The encyclopedia provides numerous contributions providing coverage of the most important issues, concepts, trends and technologies in multimedia technology each written by scholars throughout the world with notable research portfolios and expertise.

The encyclopedia also includes brief description of particular software applications or websites related to the topic of multimedia technology, networks and online interactive multimedia services.

The encyclopedia provides a compendium of terms, definitions and explanations of concepts, processes and acronyms offering an in-depth description of key terms and concepts related to different areas, issues and trends in multimedia technology and networking in modern organizations worldwide.

This encyclopedia is organized in a manner that will make your search for specific information easier and quicker. It is designed to provide thorough coverage of the field of multimedia technology and networking today by examining the following topics:

- From Circuit Switched to IP-Based Networks
  - Network Optimization
  - Information Systems in Small Firms
- Telecommunications and Networking Technologies
- Broadband Solution for the Last Mile to the Residential Customers
  - Overview
  - Copper Solutions
- Multimedia Information Management
- Mobile Computing and Commerce
  - General Trends and Economical Aspects
  - Network Evolution
- Multimedia Digital Television
- Distance Education Technologies
- Electronic Commerce Technologies Management
- End User Computing
- Information Security Management

- Open Source Technologies and Systems
- IT and Virtual Communities
- Psychology of Multimedia Technologies

The encyclopedia provides thousands of comprehensive references on existing literature and research on multimedia technologies.

In addition, a comprehensive index is included at the end of the encyclopedia to help you find cross-referenced articles easily and quickly. All articles are organized by titles and indexed by authors, making it a convenient method of reference for readers.

The encyclopedia also includes cross-referencing of key terms, figures and information related to multimedia technologies and applications.

All articles were reviewed by either the authors or by external reviewers via a blind peer-review process. In total, we were quite selective regarding inclusion of submitted articles in the encyclopedia.

INTENDED AUDIENCE

This encyclopedia will be of particular interest to teachers, researchers, scholars and professionals of the discipline, who require access to the most current information about the concepts, issues, trends and technologies in this emerging field. The encyclopedia also serves as a reference for managers, engineers, consultants, and others interested in the latest knowledge related to multimedia technology and networking.

Acknowledgements

Editing this encyclopedia was an experience without precedent, which enriched me a lot both from the human and professional side. I learned a lot from the expertise, enthusiasm, and cooperative spirit of the authors of this publication. Without their commitment to this multidisciplinary exercise, I would not have succeeded.

The efforts that we wish to acknowledge took place over the course of the last two years, as first the premises, then the project, then the challenges, and finally the encyclopedia itself took shape.

I owe a great debt to colleagues all around the world who have worked with me directly (and indirectly) on the research represented here. I am particularly indebted to all the authors involved in this encyclopedia which provided the opportunity to interact and work with the leading experts from around the world. I would like to thank all of them.

Crafting a wealth of research and ideas into a coherent encyclopedia is a process whose length and complexity I underestimated severely. I owe a great debt to Sara Reed, Assistant Managing Editor, and Renée Davies, Acquisitions/Development Editor. They helped me in organizing and carrying out the complex tasks of editorial management, deadline coordination, and page production—tasks which are normally kept separate, but which, in this encyclopedia, were integrated together so we could write and produce this book.

Mehdi Khosrow-Pour, my editor, and his colleagues at Idea Group Publishing have been extremely helpful and supportive every step of the way. Mehdi always provided encouragement and professional support. He took on this project with enthusiasm and grace, and I benefited greatly both from his working relationship with me and his editorial insights. His enthusiasm motivated me to initially accept his invitation for taking on this big project.

A further special note of thanks goes also to Jan Travers at Idea Group Publishing, whose contributions throughout the whole process from inception of the initial idea to final publication have been invaluable.

I would like to acknowledge the help of all involved in the collation and review process of the encyclopedia, without whose support the project could not have been satisfactorily completed.

Most of the authors also served as referees for articles written by other authors. Their constructive and comprehensive reviews were valuable to the overall process and quality of the final publication.

Deep appreciation and gratitude is due to members of the Editorial Advisory Board: Prof. Raymond A. Hackney of Manchester Metropolitan University (UK), Prof. Leslie Leong of Central Connecticut State University (USA), Prof. Nadia Magnenat-Thalmann of University of Geneva (Switzerland), Prof. Lorenzo Peccati of Bocconi University (Italy), Prof. Nobuyoshi Terashima of Waseda University (Japan), Prof. Steven John Simon of Mercer University (USA), Prof. Andrew Targowski of Western Michigan University (USA), Prof. Enrico Valdani of Bocconi University (Italy).

I owe a debt of gratitude to New Media&TV-lab the research laboratory on new media inside I-LAB Centre for Research on the Digital Economy of Bocconi University where I have the chance to work for the past five years. I'm deeply grateful to Prof. Enrico Valdani (Director I-LAB) for always having supported and encouraged my research endeavors inside I-LAB.

I would like to thank Prof. Charles Fine at Massachusetts Institute of Technology (Sloan School of Management) for writing the foreword of this publication. Thanks also to Anna Piccolo at Massachusetts Institute of Technology for all her support and encouragement.

Thanks go to all those who provided constructive and comprehensive reviews and editorial support services for coordination of this two year-long project.

My deepest appreciation goes to all the authors for their insights and excellent contributions to this encyclopedia. Working with them in this project was an extraordinary experience in my professional life.

In closing, I'm delighted to present this encyclopedia to you and I'm proud of the many outstanding articles that are included herein. I'm confident that you will find it to be a useful resource to help your business, your students, or your business colleagues to better understand the topics related to Multimedia Technology and Networking.

Margherita Pagani
Bocconi University
I-LAB Centre for Research on the Digital Economy
Milan, 2004

About the Editor

Dr. Margherita Pagani is head researcher for the New Media & TV-lab at the I-LAB Centre for Research on the Digital Economy of Bocconi University where she also teaches in the Management Department. She is an associate editor of the *Journal of Information Science and Technology (JIST)* and *International Journal of Cases on Electronic Commerce*. She has been a visiting scholar at the Massachusetts Institute of Technology and visiting professor at Redlands University (California). Dr. Pagani has written many refereed papers on multimedia and interactive television, digital convergence, and content management, which have been published in many academic journals and presented in academic international conferences. She has worked with Radiotelevisione Italiana (RAI) and as a member of the workgroup, "Digital Terrestrial" for the Ministry of Communications in Italy. Dr. Pagani is the author of the books "La Tv nell'era digitale" (EGEA 2000), "Multimedia and Interactive Digital TV: Managing the Opportunities Created by Digital Convergence" (IRM Press 2003), and "Full Internet mobility in a 3G-4G environment: managing new business paradigms" (EGEA 2004). She edited the Encyclopedia of Multimedia Technology and Networking (IGR 2005).

Adoption of Communication Products and the Individual Critical Mass

A

Markus Voeth
University of Hohenheim, Germany

Marcus Liehr
University of Hohenheim, Germany

THE ECONOMICS OF COMMUNICATION PRODUCTS

Communication products are characterized by the fact that the benefit that results from their use is mainly dependent on the number of users of the product, the so-called installed base, and only dependent to a minor degree on the actual product characteristics. The utility of a videoconferencing system, for example, is quite small at the product launch because only a few users are present with whom adopters can communicate. Only the increase in the number of users leads to an enhancement of the utility for each user.

The additional benefit that emerges from an increase in the installed base can be ascribed to network effects. A change in the installed base can affect the utility of products directly as well as indirectly. Direct network effects occur if the utility of a product directly depends on the number of other users of the same or a compatible product (for example, e-mail, fax machines, videoconferencing systems). Indirect network effects, on the other hand, result only indirectly from an increasing number of users because they are caused by the interdependence between the offer and demand of network products, as is the case with CD and DVD players (Katz & Shapiro, 1985). Therefore, direct network effects can be rated as demand-side network effects, while indirect network effects can be classified as supply-side network effects (Lim, Choi, & Park, 2003). For this reason, direct and indirect network effects cause different economic implications (Clements, 2004). As direct network effects predominantly appear in connection with communication products, the following observations concentrate exclusively on direct network effects.

Due to direct network effects, the diffusion of communication products is characterized by a critical mass, which "occurs at the point at which enough individuals in a system have adopted an innovation so that the innovation's further rate of adoption becomes self-sustaining" (Rogers, 2003, p. 343). Besides this market-based critical mass, there is also a critical mass at the individual level. This individual critical mass signifies the threshold of the installed base that has to be exceeded before an individual is willing to adopt a communication product (Goldenberg, Libai, & Muller, 2004).

Network effects cause a mutual dependence between the installed base and the individual willingness to adopt a communication product. This again results in the so-called start-up problem of communication products (Markus, 1987): If merely a minor installed base exists, a communication product is sufficiently attractive only for a small number of individuals who are then willing to adopt the product. However, the installed base will not increase if the communication product does not generate a sufficient utility for the potential adopters. Thus, the possibility of the failure of product diffusion is especially present at the launch of a communication product; this is due to the naturally low diffusion rate at this particular point of time and the small attractiveness resulting from this.

Therefore, the supplier of a communication product must have the aim of reaching a sufficient number of users who then continue using the product and motivate other individuals to become users, thus causing the diffusion to become self-sustaining. In this context, the management of compatibility (Ehrhardt, 2004), the timing of market entry (Srinivasan, Lilien, & Rangaswamy, 2004), penetration pricing (Lee & O'Connor, 2003), the giving away of the communication product (Shapiro & Varian, 1999), and price discrimination, which is based on the individual's social ties (Shi, 2003), are frequently discussed marketing measures. In order to market

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

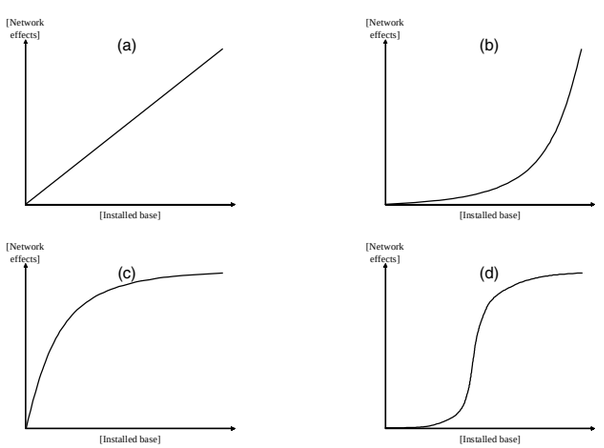
communication products, though, it is first of all necessary to gain knowledge about the characteristics of network effects and their influence on the adoption of communication products. Afterwards, the corresponding marketing measures can be derived.

CHARACTERISTICS OF NETWORK EFFECTS

Generally, two dimensions of the emergence of direct network effects are distinguished (Shy, 2001). On the one hand, network effects arise in the framework of active communication, that is, when contacting an individual in order to communicate with him or her. On the other hand, network effects also result from the possibility of being contacted by other individuals (passive communication). As direct network effects therefore result from the possibility of interacting with other users, they do not automatically arise from the purchase of a product, but rather from its use.

Regarding the functional correlation between network effects and the installed base, the literature especially distinguishes the four functional types presented in Figure 1 (Swann, 2002). While the linear function (Figure 1a) stands for the assumption that regardless of the point of time of the adoption, each new adopter causes network effects to the same degree, the convex function (Figure 1b) represents the assumption that each later adopter causes higher additional network effects than earlier adopters. Those two types of functions commonly represent the assumption that network effects are indefinitely increasing in a social system. In contrast to this, the concave and the *s*-shaped functions express the assumption that network effects are limited by a saturation level. However, while in the case of a concave function (Figure 1c) every later adopter causes lower additional network effects than earlier adopters, the *s*-shaped function (Figure 1d) is a mixture of the convex function with a low installed base and the concave function with a higher installed base. As the problem of the functional relationship between network effects and the installed base has not received much attention in the literature, there is no clear indication on the real relationship.

Figure 1. The functional form of network effects



In reference to the network effects' dependency on the number of users, An and Kiefer (1995) differentiate between network effects that depend on the worldwide installed base (global network effects) and networks effects that depend on the number of neighbouring users (local network effects). However, the abstraction from the identity of the users of a communication product often proves inadequate when practical questions are tackled (Rohlfs, 1974). As communication products serve the satisfaction of communicational needs, network effects naturally depend on the form of an individual's communication network. When deciding about the adoption of a camera cell phone, for example, people create high network effects with whom the potential adopter wants to exchange photos or videos. Therefore, it can be assumed that the adoption of people with whom the individual communicates more often or more intensively creates higher network effects than the adoption of people with a lower frequency or intensity of communication. Furthermore, groups of individuals exist, each of which display a similar communication frequency and intensity regarding the individual, thus making it necessary to differentiate between groups characterized by similarly high network effects for the individual (Voeth & Liehr, 2004).

NETWORK EFFECTS AND THE INDIVIDUAL CRITICAL MASS

Due to network effects, the adoption of communication products is characterized by the fact that the installed base has to surpass an individual threshold in order to make an individual willing to adopt the communication product. One approach at analyzing

individual thresholds is Granovetter's (1978) threshold model, which is grounded in the collective behavior literature. The aim of this model is the representation of binary decision situations, in which a rationally acting individual has to choose among two different mutually exclusive alternatives of action. For the individual, the utility and the connected costs that result from the decision here depend on the number of other individuals who have each respectively taken the same decision. In this case, the observed individual will decide on one of the two alternatives if the absolute or relative share of other individuals who have already chosen this alternative exceeds an individual threshold. When surpassing this threshold, the utility that results from the decision is at least as high as the resulting costs for the first time.

Different individuals have varying thresholds; individuals with a low threshold will thus opt for one decision alternative at a relatively early point of time, whereas individuals with a higher threshold will only decide for one alternative when a great number of other individuals have already made this decision. The distribution of individual thresholds therefore helps to explain how individual behavior influences collective behavior in a social system (Valente, 1995).

On the basis of the threshold model, the concept of an individual threshold of the installed base, which has to be surpassed in order to make an individual adopt a communication product, can be described as follows.

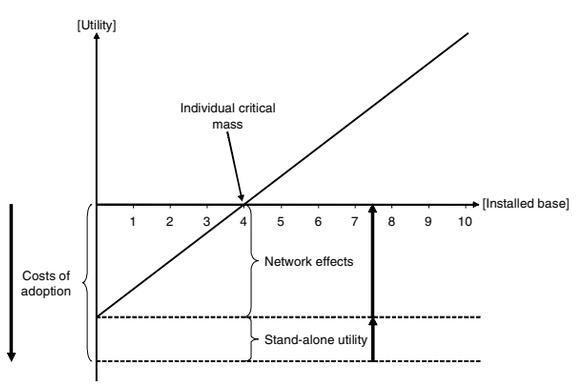
The critical threshold for the adoption of a communication product represents the degree of the installed base, for which the utility of a communication product corresponds to the costs that result from the adoption of the product; this means that the net utility of the communication product is zero and the individual is indifferent about adopting or not adopting the product. The utility of a communication product assumedly results from the sum of network effects and from the stand-alone utility that is independent from its installed base. The costs of the adoption are created mainly by the price, both of the purchase and the use of the product. Thus, the critical threshold value represents the installed base for which network effects equal the margin of the price and the stand-alone utility of a communication product. For the purpose of a notional differentiation between thresholds of collective behavior in general and thresholds of communication products in particular, this point of the installed base will hereinafter be designated as individual critical mass.

Under the simplifying assumption of a linear correlation among network effects and the installed base, the individual critical mass can be graphically determined as shown in Figure 2. In this exemplary case, the individual critical mass has the value of 4; that is, four persons have to adopt the communication product before the observed individual is willing to adopt the communication product.

The individual critical mass is the result of an individual comparison of the utility and the costs of a communication product. Therefore, it is product specific and can be influenced by a change of characteristics of the observed communication product that improve or reduce its net utility. As the individual assessments of the utility of the object characteristics vary, the individual critical masses of a certain communication product are unequally distributed in a social system. For the thresholds of collective behavior, Valente (1995) assumes that the individual thresholds are normally distributed in a social system. Due to the fact that individual critical masses can be changed by the supplier via the arrangement of the characteristics of a communication product, it is assumed for communication products that a "truncated normal distribution is a reliable working assumption" (Goldenberg et al., 2004, p. 9) "for the distribution of the individual critical mass."

The distribution of individual critical masses directly affects the diffusion of a communication product, as can be shown by the following example. If there are 10 individuals in a social system with a uniform distribution of individual critical masses ranging between 0 and 9, the diffusion process immediately starts with the adoption of the person that has an individual critical mass of 0. Subse-

Figure 2. The individual critical mass



quently, the diffusion process will continue until all members of the social system have adopted the communication product. In contrast, if the individual critical masses are continuously distributed between 1 and 10, the diffusion process will not start at all as everybody wants at least one person to adopt before they themselves do so. Against this background, a communication product should be arranged in a way that makes merely a small individual critical mass necessary for as many individuals as possible; consequently, a big share of individuals will be willing to adopt the communication product even though the installed base is small.

MEASURING INDIVIDUAL CRITICAL MASSES

The empirical measurement of thresholds in general and individual critical masses in particular have been largely neglected in the past (Lüdemann, 1999). One of the few measuring approaches of the individual critical mass was made by Goldenberg et al. (2004), who use a two-stage survey in order to determine the individual critical masses for an advanced fax machine, videoconferencing, an e-mail system, and a cell phone with picture-sending ability. As the authors intend to separate network effects from word-of-mouth effects, the informants were given a description of a scenario in which the survey object did not contain any network effects in the first step. On the basis of this scenario, the informants had to state the percentage of their friends and acquaintances who would have to adopt the survey object until they themselves would adopt it. In a second step, the authors extended the scenario by assigning network effects to the survey objects and asked the informants again to state the number of previous adopters. Because of the used scenarios, the difference of previous adopters that arises between the two stages allows a conclusion about the presence of network effects and can thus be interpreted as the individual critical mass.

The direct approach at measuring individual critical masses by Goldenberg et al. (2004) is not sufficient from a survey point of view, for the values this method arrives at are only valid for the one survey object specified in the survey. Consequently, this direct inquiry into individual critical masses has to be considered inept as a basis for the derivation of marketing activities for the following reason: If the distribution of individual critical masses is not at the core of the study, but rather the analysis of how changes in characteristics of a communication product influence individual critical masses, this extremely specific inquiry would clearly increase the time and resources required for the survey because the analysis of each change would call for the construction of a new scenario.

Against the background of this criticism, Voeth and Liehr (2004) use an indirect approach at the measuring of individual critical masses. In this approach, network effects are explicitly seen as a part of the utility of a communication product. Here, utility assessments of characteristics of communication products are asked for rather than letting informants state the number of persons who would have to have adopted the product at an earlier point of time. Subsequently, the individual critical mass can be determined as the installed base, for which the positive utility constituents of the communication product at least equal the costs of the adoption for the first time. Methodically, the measuring of individual critical masses is carried out by using a further development of the traditional conjoint analysis (Green, Krieger, & Wind, 2001), which allows conclusions about the part worth of object characteristics on the basis of holistic preference statements. In addition to the stand-alone utility components and the costs of the adoption, the installed base of a communication product is integrated into the measuring of individual critical masses as a network-effects-producing characteristic.

The chosen conjoint approach, which is designated as hierarchical limit conjoint analysis, presents a combination of limit conjoint analysis and hierarchical conjoint analysis. The limit conjoint analysis enables the integration of choice decisions into the traditional conjoint analysis and simultaneously preserves the advantage of a utility estimation on the individual level, which the traditional conjoint analysis contains (Backhaus & Voeth, 2003; Voeth, 1998). Subsequent to the traditional conjoint analysis, the informant is asked which stimuli they would be willing to buy; this makes the direct integration of choice decisions into conjoint analysis possible. The informants thus get the possibility of stating their willingness to buy one, several, all, or none of the stimuli. In

the hierarchical conjoint analysis, the object characteristics are pooled into constructs on the basis of logical considerations or empirical pilot surveys (Louviere, 1984). In a subsequent step, a conjoint design (sub design) is generated for each of these constructs, which allows the determination of the interdependence between the construct and the respective characteristics. Additionally, one more conjoint design (meta design) is generated with the constructs in order to determine the relationship between the constructs and the entire utility. The measuring of individual critical masses using the hierarchical conjoint analysis aims at the specification of the entire assessment of communication products by means of the meta design, and at the determination of the structure of network effects by means of the sub design (Voeth & Liehr, 2004). Because of this, the meta design contains the costs of the adoption of a communication product, the stand-alone utility, and network effects. The sub design, on the other hand, is used to analyze the structure of network effects by using different group-related installed bases (e.g., friends, family, acquaintances) as conjoint features.

By means of an empirical analysis of the adoption of a camera cell phone, Voeth and Liehr (2004) study the application of the hierarchical limit conjoint analysis for the measuring of individual critical masses. The main findings of this study are the following.

- As an examination of the validity of the utility estimation shows good values for face validity, internal validity, and predictive validity, the hierarchical limit conjoint analysis can be rated suitable for measuring network effects.
- The analysis of the part worth reveals that, on the one hand, the number of friends that have adopted a camera cell phone generates the highest network effects, and on the other hand, the adoptions of people the individual is not related with create only low network effects.
- In most cases, network effects tend towards a saturation level. A functional form that leads to indefinitely increasing network effects could rarely be observed.
- A high percentage of informants have an individual critical mass of zero and thus will adopt the survey product even though no one else has adopted the communication product before.
- As in the measuring approach by Goldenberg et al. (2004), the individual critical masses exhibit a bell-shaped distribution.

Although the indirect approach turns out to be suitable for the measuring of individual critical masses, continuative empirical studies regarding the suitability of different variants of conjoint analysis would be desirable.

CONCLUSION

The adoption of communication products is determined by the installed base and the network effects resulting from it. In order to derive marketing activities for communication products, it is therefore necessary to gather information about the characteristics of network effects and the level of the installed base, which is necessary in order to make an individual willing to adopt a communication product.

Based on the measurement of the individual critical mass, it is possible to determine the profitability of marketing measures for communication products. For example, if the start-up problem of a communication product is to be solved by giving away the product to selected persons in the launch period, it is not advisable to choose individuals with a low individual critical mass. Instead, it is recommendable to give the product to persons with a high individual critical mass. This is due to the fact that persons with a low individual critical mass would adopt the communication product shortly after the launch anyway, while persons with a high individual critical mass would adopt—if at all—at a much later date. Against this background, the measuring of the individual critical mass is highly relevant for the marketing of communication products.

REFERENCES

An, M. Y., & Kiefer, N. M. (1995). Local externalities and social adoption of technologies. *Journal of Evolutionary Economics, 5*(2), 103-117.

Backhaus, K., & Voeth, M. (2003). *Limit conjoint analysis* (Scientific discussion paper series no. 2). Muenster, Germany: Marketing Center Muenster, Westphalian Wilhelms University of Muenster.

Clements, M. T. (2004). Direct and indirect network effects: Are they equivalent? *International Journal of Industrial Organization, 22*(5), 633-645.

Ehrhardt, M. (2004). Network effects, standardisation and competitive strategy: How companies influence the emergence of dominant designs. *International Journal of Technology Management, 27*(2/3), 272-294.

Goldenberg, J., Libai, B., & Muller, E. (2004). *The chilling effect of network externalities on new product growth* (Working paper). Tel Aviv, Israel: Tel Aviv University.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology, 83*(6), 1420-1443.

Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces, 31*(3), S56-S73.

Katz, M. L., & Shapiro, C. (1985). Network externalities, competition, and compatibility. *American Economic Review, 75*(3), 424-440.

Lee, Y., & O'Connor, G. C. (2003). New product launch strategy for network effects products. *Journal of the Academy of Marketing Science, 31*(3), 241-255.

Lim, B.-L., Choi, M., & Park, M.-C. (2003). The late take-off phenomenon in the diffusion of telecommunication services: Network effect and the critical mass. *Information Economics and Policy, 15*(4), 537-557.

Louviere, J. J. (1984). Hierarchical information integration: A new method for the design and analysis of complex multiattribute judgment problems. *Advances in Consumer Research, 11*(1), 148-155.

Lüdemann, C. (1999). Subjective expected utility, thresholds, and recycling. *Environment and Behavior, 31*(5), 613-629.

Markus, M. L. (1987). Toward a "critical mass" theory of interactive media. *Communication Research, 14*(5), 491-511.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.

Rohlfs, J. (1974). A theory of interdependent demand for a communications service. *Bell Journal of Economics and Management Science, 5*(1), 16-37.

Shapiro, C., & Varian, H. R. (1999). *Information rules: A strategic guide to the network economy.* Boston: Harvard Business School Press.

Shi, M. (2003). Social network-based discriminatory pricing strategy. *Marketing Letters, 14*(4), 239-256.

Shy, O. (2001). *The economics of network industries.* Cambridge: Cambridge University Press.

Srinivasan, R., Lilien, G. L., & Rangaswamy, A. (2004). First in, first out? The effects of network externalities on pioneer survival. *Journal of Marketing, 68*(1), 41-58.

Swann, G. M. P. (2002). The functional form of network effects. *Information Economics and Policy, 14*(3), 417-429.

Valente, T. W. (1995). *Network models of the diffusion of innovations.* Cresskill, NJ: Hampton Press.

Voeth, M. (1998). Limit conjoint analysis: A modification of the traditional conjoint analysis. In P. Andersson (Ed.), *Proceedings of the 27th EMAC Conference, Marketing Research and Practice, Marketing Research*, (pp. 315-331).

Voeth, M., & Liehr, M. (2004). *Measuring individual critical mass and network effects* (Working paper). Hohenheim, Germany: University of Hohenheim.

KEY TERMS

Adoption: Result of an innovation decision process. Decision to use an innovation.

Conjoint Analysis: Decompositional method of preference measurement. On the basis of holistic preference statements, the part worth of object characteristics are derived.

Diffusion: Process of the spread of an innovation in a social system.

Hierarchical Conjoint Analysis: Variant of conjoint analysis that allows the integration of an extended amount of conjoint features.

Individual Critical Mass: Characteristic of the installed base that has to be surpassed in order to make an individual willing to adopt a communication product.

Installed Base: Number of current users of a certain communication product and compatible products.

Limit Conjoint Analysis: Further development of traditional conjoint analysis in which choice data is directly integrated into conjoint analysis.

Network Effects: Consumption effect in which the utility of a communication product increases with the number of other users of the same or a compatible product.

Affective Computing

Maja Pantic
Delft University of Technology, The Netherlands

INTRODUCTION

We seem to be entering an era of enhanced digital connectivity. Computers and the Internet have become so embedded in the daily fabric of people's lives that they simply cannot live without them (Hoffman et al., 2004). We use this technology to work, to communicate, to shop, to seek out new information, and to entertain ourselves. With this ever-increasing diffusion of computers in society, human-computer interaction (HCI) is becoming increasingly essential to our daily lives.

HCI design was dominated first by direct manipulation and then delegation. The tacit assumption of both styles of interaction has been that the human will be explicit, unambiguous, and fully attentive while controlling the information and command flow. Boredom, preoccupation, and stress are unthinkable, even though they are very human behaviors. This insensitivity of current HCI designs is fine for well-codified tasks. It works for making plane reservations, buying and selling stocks, and, as a matter of fact, almost everything we do with computers today. But this kind of categorical computing is inappropriate for design, debate, and deliberation. In fact, it is the major impediment to having flexible machines capable of adapting to their users and their level of attention, preferences, moods, and intentions.

The ability to detect and understand affective states of a person with whom we are communicating is the core of emotional intelligence. Emotional intelligence (EQ) is a facet of human intelligence that has been argued to be indispensable and even the most important for a successful social life (Goleman, 1995). When it comes to computers, however, not all of them will need emotional intelligence, and none will need all of the related skills that we need. Yet man-machine interactive systems capable of sensing stress, inattention, and heedfulness, and capable of adapting and responding appropriately to these affective states of the user are likely to be perceived as more natural, more efficacious and more trustworthy. The research area of machine analysis and employment of human affective states to build more natural, flexible HCI goes by a general name of affective computing, introduced first by Picard (1997).

BACKGROUND: RESEARCH MOTIVATION

Besides the research on natural, flexible HCI, various research areas and technologies would benefit from efforts to model human perception of affective feedback computationally. For instance, automatic recognition of human affective states is an important research topic for video surveillance as well. Automatic assessment of boredom, inattention, and stress will be highly valuable in situations where firm attention to a crucial but perhaps tedious task is essential, such as aircraft control, air traffic control, nuclear power plant surveillance, or simply driving a ground vehicle like a truck, train, or car. An automated tool could provide prompts for better performance, based on the sensed user's affective states.

Another area that would benefit from efforts toward computer analysis of human affective feedback is the automatic affect-based indexing of digital visual material. A mechanism for detecting scenes or frames that contain expressions of pain, rage, and fear could provide a valuable tool for violent-content-based indexing of movies, video material, and digital libraries.

Other areas where machine tools for analysis of human affective feedback could expand and enhance research and applications include specialized areas in professional and scientific sectors. Monitoring and interpreting affective behavioral cues are important to lawyers, police, and security agents who are often interested in issues concerning deception and attitude. Machine analysis of human affec-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. The main problem areas in the research on affective computing

• *What is an affective state?* This question is related to psychological issues pertaining to the nature of affective states and the way affective states are to be described by an automatic analyzer of human affective states. • *What kinds of evidence warrant conclusions about affective states?* In other words, which human communicative signals convey messages about an affective arousal? This issue shapes the choice of different modalities to be integrated into an automatic analyzer of affective feedback. • *How can various kinds of evidence be combined to generate conclusions about affective states?* This question is related to neurological issues of human sensory-information fusion, which shape the way multi-sensory data is to be combined within an automatic analyzer of affective states.

tive states could be of considerable value in these situations where only informal interpretations are now used. It would also facilitate research in areas such as behavioral science (in studies on emotion and cognition), anthropology (in studies on cross-cultural perception and production of affective states), neurology (in studies on dependence between emotional abilities impairments and brain lesions), and psychiatry (in studies on schizophrenia) in which reliability, sensitivity, and precision are persisting problems.

BACKGROUND: THE PROBLEM DOMAIN

While all agree that machine sensing and interpretation of human affective information would be quite beneficial for manifold research and application areas, addressing these problems is not an easy task. The main problem areas are listed in Table 1.

On one hand, classic psychological research follows from the work of Darwin and claims the existence of six basic expressions of emotions that are universally displayed and recognized: happiness, anger, sadness, surprise, disgust, and fear (Lewis & Haviland-Jones, 2000). In other words, all non-verbal communicative signals (i.e., facial expression, vocal intonations, and physiological reactions) involved in these basic emotions are displayed and recognized cross-culturally. On the other hand, there is now a growing body of psychological research that strongly challenges the classical theory on emotion. Russell (1994) argues that emotion in general can best be characterized in terms of a multi-dimensional affect space, rather than in terms of a small number of emotion categories. Social constructivists argue that emotions are socially constructed ways of interpreting and responding to particular classes of situations and that they do not explain the genuine feeling (affect). Also, there is no consensus on how affective displays should be labeled (Wierzbicka, 1993). The main issue here is that of culture dependency; the comprehension of a given emotion label and the expression of the related emotion seem to be culture dependent (Matsumoto, 1990). In summary, it is not certain that each of us will express a particular affective state by modulating the same communicative signals in the same way, nor is it certain that a particular modulation of interactive cues will be interpreted always in the same way independent of the situation and the observer. The immediate implication is that pragmatic choices (e.g., application- and user-profiled choices) must be made regarding the selection of affective states to be recognized by an automatic analyzer of human affective feedback.

Affective arousal modulates all verbal and non-verbal communicative signals (Ekman & Friesen, 1969). Hence, one could expect that automated human-affect analyzers should include all human interactive modalities (sight, sound, and touch) and should analyze all non-verbal interactive signals (facial expressions, vocal expressions, body gestures, and physiological reactions). Yet the reported research does not confirm this assumption. The visual channel carrying facial expressions and the auditory channel carrying vocal intonations are widely thought of as most important in the human recognition of affective feedback. According to Mehrabian

Table 2. The characteristics of an ideal automatic human-affect analyzer

- **multimodal** (modalities: facial expressions, vocal intonations)
- **robust and accurate** (despite auditory noise, occlusions and changes in viewing and lighting conditions)
- **generic** (independent of variability in subjects' physiognomy, sex, age and ethnicity)
- **sensitive to the dynamics** (time evolution) of displayed affective expressions (performing temporal analysis of the sensed data, previously processed in a joint feature space)
- **context-sensitive** (performing application- and task-dependent data interpretation in terms of user-profiled affect-interpretation labels)

(1968), whether the listener feels liked or disliked depends on 7% of the spoken word, 38% on vocal utterances, and 55% on facial expressions. This indicates that while judging someone's affective state, people rely less on body gestures and physiological reactions displayed by the observed person; they rely mainly on facial expressions and vocal intonations. Hence, automated affect analyzers should at least combine modalities for perceiving facial and vocal expressions of affective states.

Humans simultaneously employ the tightly coupled modalities of sight, sound, and touch. As a result, analysis of the perceived information is highly robust and flexible. Hence, in order to accomplish a multimodal analysis of human interactive signals acquired by multiple sensors, which resembles human processing of such information, input signals cannot be considered mutually independent and cannot be combined only at the end of the intended analysis, as the majority of current studies do. The input data should be processed in a joint feature space and according to a context-dependent model (Pantic & Rothkrantz, 2003).

In summary, an ideal automatic analyzer of human affective information should be able to emulate at least some of the capabilities of the human sensory system (Table 2).

THE STATE OF THE ART

Facial expressions are our primary means of communicating emotion (Lewis & Haviland-Jones, 2000), and it is not surprising, therefore, that the majority of efforts in affective computing concern automatic analysis of facial displays. For an exhaustive survey of studies on machine analysis of facial affect, the readers are referred to Pantic and Rothkrantz (2003). This survey indicates that the capabilities of currently existing facial affect analyzers are rather limited (Table 3). Yet, given that humans detect six basic emotional facial expressions with an accuracy ranging from 70% to 98%, it is rather significant that the automated systems achieve an accuracy of 64% to 98% when detecting three to seven emotions deliberately displayed by five to 40 sub-

Table 3. Characteristics of currently existing automatic facial affect analyzers

- **handle a small set of posed prototypic facial expressions** of six basic emotions from portraits or nearly-frontal views of faces with no facial hair or glasses recorded under good illumination
- **do not perform a task-dependent interpretation** of shown facial behavior – yet, a shown facial expression may be misinterpreted if the current task of the user is not taken into account (e.g., a frown may be displayed by the speaker to emphasize the difficulty of the currently discussed problem and it may be shown by the listener to denote that he did not understand the problem at issue)
- **do not analyze extracted facial information on different time scales** (proposed inter-video-frame analyses are usually used to handle the problem of partial data) – consequently, automatic recognition of the expressed mood and attitude (longer time scales) is still not within the range of current facial affect analyzers

jects. An interesting point, nevertheless, is that we cannot conclude that a system achieving a 92% average recognition rate performs better than a system attaining a 74% average recognition rate when detecting six basic emotions from face images. Namely, in spite of repeated references to the need for a readily accessible reference set of images (image sequences) that could provide a basis for benchmarks for efforts in automatic facial affect analysis, no database of images exists that is shared by all diverse facial-expression-research communities.

If we consider the verbal part (strings of words) only, without regard to the manner in which it was spoken, we might miss important aspects of the pertinent utterance and even misunderstand the spoken message by not attending to the non-verbal aspect of the speech. Yet, in contrast to spoken language processing, which has witnessed significant advances in the last decade, vocal expression analysis has not been widely explored by the auditory research community. For a survey of studies on automatic analysis of vocal affect, the readers are referred to Pantic and Rothkrantz (2003). This survey indicates that the existing automated systems for auditory analysis of human affect are quite limited (Table 4). Yet humans can recognize emotion in a neutral-content speech with an accuracy of 55% to 70% when choosing from among six basic emotions, and automated vocal affect analyzers match this accuracy when recognizing two to eight emotions deliberately expressed by subjects recorded while pronouncing sentences having a length of one to 12 words. Similar to the case of automatic facial affect analysis, no readily accessible reference set of speech material exists that could provide a basis for benchmarks for efforts in automatic vocal affect analysis.

Relatively few of the existing works combine different modalities into a single system for human affective state analysis. Examples are the works of Chen and Huang (2000), De Silva and Ng (2000), Yoshitomi et al. (2000), Go et al. (2003), and Song et al. (2004), who investigated the effects of a combined detection of facial and vocal expressions of affective states. In brief, these studies assume clean audiovisual input (e.g., noise-free recordings, closely-placed microphone, non-occluded portraits) from an actor speaking a single word and displaying exaggerated facial expressions of a basic emotion. Though audio and image processing techniques in these systems are relevant to the discussion on the state of the art in affective computing, the systems themselves have all (as well as some additional) drawbacks of single-modal affect analyzers and, in turn, need many improvements, if they are to be used for a multimodal context-sensitive HCI, where a clean input from a known actor/announcer cannot be expected and a context-independent data interpretation does not suffice.

CRITICAL ISSUES

Probably the most remarkable issue about the state of the art in affective computing is that, although the

Table 4. Characteristics of currently existing automatic vocal affect analyzers

- **perform singular classification of input audio signals into a few emotion categories** such as anger, irony, happiness, sadness/grief, fear, disgust, surprise and affection
- do not perform a context-sensitive analysis (i.e., application-, user- and task-dependent analysis) of the input audio signal
- **do not analyze extracted vocal expression information on different time scales** (proposed inter-audio-frame analyses are used either for the detection of supra-segmental features, such as the pitch and intensity over the duration of a syllable, word, or sentence, or for the detection of phonetic features) – computer-based recognition of moods and attitudes (longer time scales) from input audio signal remains a significant research challenge
- **adopt strong assumptions to make the problem of automating vocal-expression analysis more tractable** (e.g., the recordings are noise free, the recorded sentences are short, delimited by pauses, carefully pronounced by non-smoking actors to express the required affective state) and use the test data sets that are small (one or more words or one or more short sentences spoken by few subjects) containing exaggerated vocal expressions of affective states

recent advances in video and audio processing make audiovisual analysis of human affective feedback tractable, and although all agreed that solving this problem would be extremely useful, merely a couple of efforts toward the implementation of such a bimodal human-affect analyzer have been reported to date.

Another issue concerns the interpretation of audiovisual cues in terms of affective states. The existing work employs usually singular classification of input data into one of the basic emotion categories. However, pure expressions of basic emotions are seldom elicited; most of the time, people show blends of emotional displays. Hence, the classification of human non-verbal affective feedback into a single basic-emotion category is not realistic. Also, not all non-verbal affective cues can be classified as a combination of the basic emotion categories. Think, for instance, about the frustration, stress, skepticism, or boredom. Furthermore, it has been shown that the comprehension of a given emotion label and the ways of expressing the related affective state may differ from culture to culture and even from person to person. Hence, the definition of interpretation categories in which any facial and/or vocal affective behavior, displayed at any time scale, can be classified is a key challenge in the design of realistic affect-sensitive monitoring tools. One source of help is machine learning; the system potentially can learn its own expertise by allowing the user to define his or her own interpretation categories (Pantic, 2001).

Accomplishment of a human-like interpretation of sensed human affective feedback requires pragmatic choices (i.e., application-, user- and task-profiled choices). Nonetheless, currently existing methods aimed at the automation of human-affect analysis are not context sensitive. Although machine-context sensing (i.e., answering questions like who is the user, where is the user, and what is the user doing) has witnessed recently a number of significant advances (Pentland, 2000), the complexity of this problem makes context-sensitive human-affect analysis a significant research challenge.

Finally, no readily accessible database of test material that could be used as a basis for benchmarks for efforts in the research area of automated human affect analysis has been established yet. In fact, even in the research on facial affect analysis, which attracted the interest of many researchers, there is a glaring lack of an existing benchmark face database. This lack of common testing resources forms the major impediment to comparing, resolving, and extending the issues concerned with automatic human affect analysis and understanding. It is, therefore, the most critical issue in the research on affective computing.

CONCLUSION

As remarked by scientists like Pentland (2000) and Oviatt (2003), multimodal context-sensitive (user-, task-, and application-profiled and affect-sensitive) HCI is likely to become the singlemost widespread research topic of the AI research community. Breakthroughs in such HCI designs could bring about the most radical change in the computing world; they could change not only how professionals practice computing, but also how mass consumers conceive and interact with the technology. However, many aspects of this new-generation HCI technology, in particular ones concerned with the interpretation of human behavior at a deeper level and the provision of the appropriate response, are not mature yet and need many improvements.

REFERENCES

Chen, L.S., & Huang, T.S. (2000). Emotional expressions in audiovisual human computer interaction. *Proceedings of the International Conference on Multimedia and Expo.,* New York, (pp. 423-426).

De Silva, L.C., & Ng, P.C. (2000). Bimodal emotion recognition. *Proceedings of the. International Conference on Face and Gesture Recognition*, Grenoble, France, (pp. 332-335).

Ekman, P., & Friesen, W.F. (1969). The repertoire of nonverbal behavioral categories—Origins, usage, and coding. *Semiotica, 1*, 49-98.

Go, H.J., Kwak, K.C., Lee, D.J., & Chun, M.G. (2003). Emotion recognition from facial image and speech signal. *Proceedings of the Conference of*

the Society of Instrument and Control Engineers, Fukui, Japan, (pp. 2890-2895).

Goleman, D. (1995). *Emotional intelligence*. New York: Bantam Books.

Hoffman, D.L., Novak, T.P., & Venkatesh, A. (2004). Has the Internet become indispensable? *Communications of the ACM, 47*(7), 37-42.

Lewis, M., & Haviland-Jones, J.M. (Eds.). (2000). *Handbook of emotions*. New York: Guilford Press.

Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motivation and Emotion, 14*, 195-214.

Mehrabian, A. (1968). Communication without words. *Psychology Today, 2*(4), 53-56.

Oviatt, S. (2003). User-centered modeling and evaluation of multimodal interfaces. *Proceedings of the IEEE, 91*(9), 1457-1468.

Pantic, M. (2001). *Facial expression analysis by computational intelligence techniques* [Ph.D. Thesis]. Delft, Netherlands: Delft University of Technology.

Pantic, M., & Rothkrantz, L.J.M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE, 9*1(9), 1370-1390.

Pentland, A. (2000). Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 107-119.

Picard, R.W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Russell, J.A. (1994). Is there universal recognition of emotion from facial expression? *Psychological Bulletin, 115*(1), 102-141.

Song, M., Bu, J., Chen, C., & Li, N. (2004). Audio-visual based emotion recognition—A new approach. *Proceedings of the International Conference Computer Vision and Pattern Recognition*, Washington, USA, (pp. 1020-1025).

Wierzbicka, A. (1993). Reading human faces. *Pragmatics and Cognition, 1*(1), 1-23.

Yoshitomi, Y., Kim, S., Kawano, T., & Kitazoe, T. (2000). Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. *Proceedings of the International Workshop on Robot-Human Interaction*, Osaka, Japan, (pp. 178-183).

A

KEY TERMS

Affective Computing: The research area concerned with computing that relates to, arises from, or deliberately influences emotion. Affective computing expands HCI by including emotional communication, together with the appropriate means of handling affective information.

Benchmark Audiovisual Affect Database: A readily accessible centralized repository for retrieval and exchange of audio and/or visual training and testing material and for maintaining various test results obtained for a reference audio/visual data set in the research on automatic human affect analysis.

Context-Sensitive HCI: HCI in which the computer's context with respect to nearby humans (i.e., who the current user is, where the user is, what the user's current task is, and how the user feels) is automatically sensed, interpreted, and used to enable the computer to act or respond appropriately.

Emotional Intelligence: A facet of human intelligence that includes the ability to have, express, recognize, and regulate affective states, employ them for constructive purposes, and skillfully handle the affective arousal of others. The skills of emotional intelligence have been argued to be a better predictor than IQ for measuring aspects of success in life.

Human-Computer Interaction (HCI): The command and information flow that streams between the user and the computer. It is usually characterized in terms of speed, reliability, consistency, portability, naturalness, and users' subjective satisfaction.

Human-Computer Interface: A software application, a system that realizes human-computer interaction.

Multimodal (Natural) HCI: HCI in which command and information flow exchanges via multiple natural sensory modes of sight, sound, and touch. The user commands are issued by means of speech, hand gestures, gaze direction, facial expressions, and so forth, and the requested information or the computer's feedback is provided by means of animated characters and appropriate media.

Agent Frameworks

Reinier Zwitserloot
Delft University of Technology, The Netherlands

Maja Pantic
Delft University of Technology, The Netherlands

INTRODUCTION

Software agent technology generally is defined as the area that deals with writing software in such a way that it is autonomous. In this definition, the word *autonomous* indicates that the software has the ability to react to changes in its environment in a way that it can continue to perform its intended job. Specifically, changes in its input channels, its output channels, and the changes in or the addition or removal of other agent software should cause the agent to change its own behavior in order to function properly in the new environment. In other words, the term *software agent* refers to the fact that a certain piece of software likely will be able to run more reliably without user intervention in a changing environment compared to similar software designed without the software agent paradigm in mind. This definition is quite broad; for example, an alarm clock that automatically accounts for daylight savings time could be said to be autonomous in this property; a change in its environment (namely, the arrival of daylight savings time) causes the software running the clock to adjust the time it displays to the user by one hour, preserving, in the process, its intended function—displaying the current time. A more detailed description of agent technology is available from Russel and Norvig (2003).

The autonomous nature of software agents makes them the perfect candidate for operating in an environment where the available software continually changes. Generally, this type of technology is referred to as multi-agent systems (MAS). In the case of MAS, the various agents running on the system adapt and account for the other agents available in the system that are relevant to its own operation in some way. For example, MAS-aware agents often are envisioned to have a way of negotiating for the use of a scarce resource with other agents.

An obvious start for developing MAS is to decide on a common set of rules to which each agent will adhere, and on an appropriate communication standard. These requirements force the need for an underlying piece of software called an agent framework. This framework hosts the agents, is responsible for ensuring that the agents keep to the rules that apply to the situation, and streamlines communication between the agents themselves and external sensors and actuators (in essence, input and output, respectively). This paper will go into more detail regarding the advantages of MAS and agent frameworks, the nature and properties of agent frameworks, a selection of frameworks available at the moment, and attempts to draw some conclusions and best practices by analyzing the currently available framework technology.

BACKGROUND: RESEARCH MOTIVATIONS

An agent framework and its use as a base for MAS technology already has been successfully used as the underlying technology for most teams participating in the robot soccer tournament (Tambe, 1998). The robotic soccer tournament requires that all participating robot teams operate entirely under their own control without any intervention by their owners. The general idea of independent autonomous robots working together to perform a common task can be useful in many critical situations. For example, in rescue situations, a swarm of heterogeneous (not the same hardware and/or software) agents controlling various pieces of hardware fitted onto robots potentially can seek out and even rescue

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

people trapped in a collapsed building. The ideal strived for in this situation is a system whereby a number of locator robots, equipped with a legged transport system to climb across any obstacle and sporting various location equipment such as audio and heat sensors, will rapidly traverse the entirety of the disaster area, creating a picture of potential rescue sites. These, in turn, serve as the basis for heavy tracked robots equipped with digging equipment, which work together with structure scanning robots that help the digging robots decide which pieces to move in order to minimize the chances of accidentally causing a further collapse in an unstable pile of rubble. Equipment breaking down or becoming disabled, for example, due to getting crushed under an avalanche of falling rubble, or falling down in such a way that it can't get up, are not a problem when such a rescue system is designed with MAS concepts in mind; as all agents (each agent powering a single robot in the system) are independent and will adapt to work together with other robots that currently are still able to operate, there is no single source of system failure, which is the case when there is a central computer controlling the system. Another advantage of not needing a central server is the ability to operate underground or in faraway places without a continuous radio link, which can be difficult under the previously mentioned circumstances.

A crucial part of such a redundancy-based system, where there are no single sources of failure, is to have backup sensor equipment. In the case of conflicts between separate sensor readings that should have matched, agents can negotiate among themselves to decide on the action to take to resolve the discrepancy. For example, if a teacup falls to the floor, and the audio sensor is broken, the fact that the video and image processing equipment registered the fall of the teacup will result in a negotiation session. The teacup fell according to the agent controlling video analysis, but the audio analyzer determined that the teacup did not fall—there was no sound of the shattering cup. In these cases, the two agents most likely will conclude the teacup did fall in the end, especially if the audio agent is capable of realizing something may be wrong with its sensors due to the video backup. Or the agents together can determine if further detail is required and ask an agent in control of a small reconnaissance robot to move to the projected site where the teacup fell and inspect the floor for cup fragments. The system will still be able to determine the need to order new teacups, even though the audio sensor that usually determines the need for new teacups currently is broken. This example displays one of the primary research motivations for multi-agent systems and agent frameworks—the ability to continue operation even if parts of the system are damaged or unavailable. This aspect is in sharp contrast to the usual state of affairs in the world of computer science; for example, even changing a single bit in a stream of code of a word processor program usually will break it to the point that it will not function at all.

Another generally less important but still significant motivation for MAS research is the potential benefit of using it as a basis for systems that exhibit emergent behavior. Emergent behavior refers to complex behavior of a system of many agents, even though none of the individual components (agents) has any kind of complex code. Emergent behavior is functionally equivalent to the relatively complex workings of a colony of ants capable of feeding the colony, relocating the hive when needed, and fending off predators, even though a single ant is not endowed at all with any kind of advanced brain function. More specifically, ants always will dispose of dead ants at the point that is farthest away from all colony entrances. A single ant clearly cannot solve this relatively complex geometrical problem; even a human being needs mathematical training before being able to solve such a geometric problem. The ability to find the answer to the problem of finding the farthest point from a set of points is an emergent ability displayed by ant colonies. The goal of emergent behavior research is to create systems that are robust in doing a very complex job, even with very simple equipment, contrasted to products that are clunky to use, hard to maintain, and require expensive equipment, as created by traditional programming styles. Areas where emergent behavior has proven to work can be found first and foremost in nature: Intelligence is evidently an emergent property; a single brain cell is government by extremely simple rules, whereas a brain is the most complex computer system known to humankind. This example also highlights the main problem with emergent behavior research; predicting what, if any, emergent behavior will occur is almost impossible.

Conversely, figuring out why a certain observed emergent behavior occurs, given the rules of the base component, usually is not an easily solved problem. While the neuron is understood, the way a human brain functions is not. Still, research done so far is promising. The most successes in this area are being made by trying to emulate emergent behavior observed in nature. Bourjot (2003) provides an example of this phenomenon. These promising results also are motivating agent framework research in order to improve the speed and abilities of the underlying building blocks of emergent behavior research—simple agents operating in an environment with many such simple agents.

PROPERTIES OF AGENT FRAMEWORKS

Many different philosophies exist regarding the design of an agent framework. As such, standardization attempts such as MASIF, KQML, and FIPA mostly restrict themselves to some very basic principles, unfortunately resulting in the virtual requirement to offer features that exceed the specification of the standard. Possibly, this aspect is the main reason that standards adherence is not common among the various agent frameworks available. Instead, a lot of frameworks appear to be developed with a very specific goal in mind. As can be expected, these frameworks do very well for their specific intended purpose. For example, hive focuses on running large amounts of homogenous (i.e., all agents have the same code) agents as a way to research emergent behavior and is very useful in that aspect. This section analyzes the basic properties of the various agent frameworks that are currently available.

- **Programming Language:** Implementing the agent will require writing code or otherwise instructing the framework on how to run the agent. Hence, one of the first things noted when inspecting an agent framework is which language(s) can be used. A lot of frameworks use Java, using the write-once-run-anywhere philosophy of the language designers to accentuate the adaptable nature of agent software. However, *C++*, Python, and a language specification specialized for creating distributed software called *CORBA* also are available. Some frameworks take a more specific approach and define their own language or present some sort of graphical building tool as a primary method of defining agent behavior (e.g., ZEUS). A few frameworks (e.g., MadKit) even offer a selection of languages. Aside from the particulars of a potential agent author, the programming language can have a marked effect on the operation of the framework. For example, *C++* based frameworks tend not to have the ability to prevent an agent from hogging system resources due to the way natively compiled code (such as that produced by a *C++* compiler) operates. Java programs inherently can be run on many different systems, and, as a result, most Java-based frameworks are largely OS and hardware independent. Frameworks based on CORBA result in a framework that has virtually no control or support for the agent code but is very flexible in regard to programming language. Due to the highly desirable properties of system independence offered by the Java programming language, all frameworks reviewed in the next section will be based on the Java language.
- **State Saving and Mobility:** The combination of the autonomous and multi-agent paradigm results in a significant lowering of the barrier for distributed computing. The agent software is already written to be less particular about the environment in which it is run, opening the door for sending a running agent to another computer. Multi-agent systems themselves also help in realizing distributed computing. An agent about to travel to another system can leave a copy of itself behind to facilitate communication of its actions on the new system back to its place of origin. As a result, a lot of agent frameworks offer the ability to move to another host to its agents (e.g., Fleeble, IBM Aglets, NOMADS, Voyager, Grasshopper). The ability to travel to other hosts is called mobility. Advantages of mobility include the ability of code, which is relatively small, to move to a large volume of data, thus saving significant bandwidth. Another major advantage is the ability to use computer resources (i.e., memory, CPU) that

are not otherwise being used on another computer—the creation of a virtual mega computer by way of combining the resources of many ordinary desktop machines. Inherent in the ability to move agents is the ability to save the state of an agent. This action freezesthe agent and stores all relevant information (the state). This stored state then either can be restored at a later time or, alternatively, can be sent to another computer to let it resume running the agent (mobility). The difficulty in true mobility lies in the fact that it is usually very difficult to just interrupt a program while it is processing. For example, if an agent is accessing a file on disk while it is moved, the agent loses access to the file in the middle of an operation. Demanding from the agent framework that it check in with the framework often, in a state where it is not accessing any local resources that cannot be moved along with the agent, generally solves this problem (Tryllian).

- **Communication Strategy:** There are various communication strategies used by frameworks to let agents talk to each other and to sensors and actuators. A common but hard-to-scale method is the so-called multicast strategy, which basically connects all agents on the system to all other agents. In the multicast system, each agent is responsible for scanning all incoming communications for whether or not an agent should act or account for the data. A more refined version of the multicast strategy is the publish/subscribe paradigm. In this system, agents can create a chat room, usually called a channel, and publish information to it, in the form of messages. Only those agents that have been subscribed to a particular channel will receive the *messages*. This solution is more flexible, especially when the framework hosts many agents. Other, less frequent strategies include *a direct* communication where data can only be sent to specific agents, or, for some systems, no communication ability exists at all.
- **Resource Management:** Exhausting the local system's processing power and memory resources is a significant risk when running many agents on one system, which, by definition, run continuously and all at the same time. Some frameworks take control of distributing the available system resources (i.e., memory, CPU, disk space, etc) and will either preventively shut down agents using too many resources or simply deny access to them. Unfortunately, the frequent monitoring of the system required to schedule the available resources results in a fairly significant CPU overhead and sometimes impedes the flexibility of the framework. For example, NOMADS uses a modified version of the Java runtime environment to implement its monitoring policy, unfortunately causing NOMADS to be out of date, compared to Sun's current version of the Java runtime environment at the time of writing. While many frameworks choose to forego resource management for these reasons, a framework that supports resource management can create a true sandbox for its agents, a place where the agent cannot harm or impact the host computer in any way, thus allowing the safe execution of agents whose code is not currently trusted. Such a sandbox system can enable the ability to run agents shared by other people, even if you don't particularly trust that their systems are free of viruses, for example. In addition to CPU and memory resource management, a proper sandbox system also needs to restrict and monitor access to data sources, such as access to the network and system storage, such as a hard drive. Some programming languages, including Java, have native support for this kind of security measure. As a result, some frameworks exist that implement this aspect of agent frameworks (SeMoA). By itself, this limited form of resource management will prevent direct damage to the local system by, for example, using the computer's network connection to attack a Web site, but can't stop an agent from disabling the host system. Due to the nature of *C++*, no C++ based frameworks support any kind of resource management.

THE STATE OF THE ART

Table 1 summarizes the properties of the currently available Java-based agent frameworks with respect to the following issues (Pantic et al., 2004):

1. Does the developer provide support for the tool?
2. Is the tool available for free?
3. Are useful examples readily available?
4. Is the related documentation readable?
5. Is synchronous agent-to-agent communication (i.e., wait for reply) supported?
6. Is asynchronous agent-to-agent communication (continuing immediately) supported?
7. What is the communication transmission form?
8. Can the framework control agents' resources (e.g., disk or network capacity used)?
9. Can the framework ask an agent to shut down?
10. Can the framework terminate the execution of a malfunctioning agent?
11. Can the framework store agents' states between executions?
12. Can the framework store objects (e.g., a database) between executions?
13. Does a self-explicatory GUI per agent exist?
14. Does the GUI support an overview of all running agents?

A detailed description of agent frameworks 1-5, 7, 9-18, and 20-24 can be found at AgentLink (2004). A detailed description of CIAgent framework is given by Bigus and Bigus (2001). More information on FIPA-OS is available at Emorphia Research (2004). Pathwalker information is provided by Fujitsu Labs (2000). More information on Tagents can be found at IEEE Distributed Systems (2004). Information on the Fleeble Framework is available from Pantic et al. (2004). The chart shows the emergence of certain trends. For example, termination of malfunctioning agents (i.e.: those that take too many or restricted resources) is offered by only a very small number of frameworks, as shown by columns 8 and 10. Another unfortunate conclusion that can be made from columns 3 and 4 is the lack of proper documentation for most frameworks. The learning curve for such frameworks is needlessly high and seems to be a factor contributing to the large selection of frameworks available.

Sharing a framework so that it is used in as many places as possible has many advantages due to the nature of a framework; namely, to serve as a standard for which agents can be written. Hence, a simple learning curve, supported by plenty of examples and good documentation is even more important than is usual in the IT sector.

Table 1. Overview of the available Java-based agent frameworks

Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Agent Factory	●	-	×	×	×	×	PP	×	×	×	×	×	×	×
2. IBM Aglets	●	Free	×	×	●	●	PP	●	●	×	×	●	●	×
3. AMETAS	●	-	×	×	×	●	M	×	●	×	×	×	●	×
4. Beegent	×	-	●	×	×	●	PP	×	×	×	×	-	×	×
5. Cougaar	●	Free	×	×	×	●	M	●	×	×	×	●	×	×
6. CIAgent	●	$	●	●	×	●	PP	×	●	×	×	-	●	×
7. DECAF	●	-	×	×	×	●	PP	×	×	×	×	×	●	×
8. FIPA-OS	●	Free	×	×	●	●	PP	×	×	×	×	●	×	×
9. Grasshopper	●	-	●	×	●	●	M	●	-	-	●	×	●	●
10. Hive	×	Free	×	×	×	×	None	×	×	×	×	×	●	●
11. JACK	●	$	×	×	×	●	PP	×	×	×	×	-	×	×
12. JADE	●	Free	●	×	×	●	M	×	●	×	×	×	●	●
13. JAFMAS / Jive	×	-	×	×	●	×	M	×	×	×	×	×	×	×
14. Kaariboga	●	Free	●	●	×	●	PP	●	●	×	×	×	×	×
15. LIME	●	Free	●	●	×	●	PS	×	×	×	×	×	×	×
16. MadKit	●	Free	●	×	●	●	M	×	●	×	×	×	×	●
17. NOMADS	×	Free	×	×	×	×	None	●	●	●	×	×	●	×
18. OpenCybele	●	Free	●	●	●	●	PS	×	●	●	×	×	×	×
19. Pathwalker	●	Free	×	×	×	●	PP	×	×	×	×	×	×	×
20. SeMoA	●	Free	●	●	×	●	None	●	●	●	●	●	×	×
21. Tagent	×	Free	×	×	●	●	PP	×	×	×	×	×	×	×
22. Tryllian	●	$	●	×	●	×	PP	×	-	-	●	●	×	×
23. Voyager	●	$	×	×	●	●	PS	●	×	×	●	×	×	×
24. ZEUS	×	Free	●	×	●	×	PP	×	×	×	×	×	×	×
25. Fleeble	●	Free	●	●	×	●	PS	●	×	●	●	●	●	●

Legend: ● = "yes", × = "no", - = unknown, PP = Peer to Peer, M = Multicast, PS = Publish -Subscribe

FUTURE TRENDS: SIMPLICITY

Fulfilling the MAS ideal of creating a truly adaptive, autonomous agent is currently impeded by steep learning curves and lack of flexibility in the available frameworks. Hence, a promising new direction for the agent framework area is the drive for simplicity, which serves the dual purpose of keeping the software flexible while making it relatively simple to write agents for the framework. Newer frameworks such as Fleeble forego specialization to try to attain this ideal. The existence of emergent behavior proves that simplistic agents are still capable of being used to achieve very complex results. Frameworks that give its agents only a limited but flexible set of commands while rigidly enforcing the MAS ideal that one agent cannot directly influence another enables the use of such a framework in a very wide application domain, from a control platform for a swarm of robots to a software engineering paradigm to reduce bugs in complex software by increasing the level of independence between parts of the software, thereby offering easier and more robust testing opportunities. Another area in which simplicity is inherently a desirable property is the field of education. The ability to let agents representing the professor or teacher inspect and query agents written to complete assignments by students represents a significant source of time-saving, enabling adding more hands-on practical work to the curriculum. A framework that is simple to use and understand is a requirement for basing the practical side of CS education on writing agents. More information on using agent frameworks as a teaching tool is available from Pantic (2003).

CONCLUSION

Agent framework technology lies at the heart of the multi-agent systems branch of artificial iIntelligence. While many frameworks are available, most differ substantially in supported programming languages, ability to enable agents to travel (mobility), level of resource management, and the type of communication between agents that the framework supports. Emergent behavior, a research area focusing on trying to create complex systems by letting many simple agents interact, along with a need for flexibility, is driving research toward providing more robust and less complex frameworks.

REFERENCES

AgentLink (2004*). http://www.agentlink.org/resources/agent-software.php*

Bigus, J.P., & Bigus J. (2001). *Constructing intelligent agent using Java.* Hoboken, NJ: Wiley & Sons.

Bourjot, C., Chevrier, V., & Thomas, V. (2003). A new swarm mechanism based on social spiders colonies: From Web weaving to region detection. *Web Intelligence and Agent Systems: An International Journal, 1*(1), 47-64.

Emorphia Research. (2004). *http://www.emorphia.com/research/about.htm*

Fujitsu Labs. (2000). *http://www.labs.fujitsu.com/en/freesoft/paw/*

IEEE Distributed Systems. (2004). *http://dsonline.computer.org/agents/projects.htm*

Pantic, M., Zwitserloot, R., & Grootjans, R.J. (2003, August). Simple agent framework: An educational tool introducing the basics of AI programming. *Proceedings of the IEEE International Conference on Information Technology: Research and Education (ITRE '03)*, Newark, USA.

Pantic, M., Zwitserloot, R., & Grootjans, R.J. (2004). Teaching introductory artificial intelligence using a simple agent framework [accepted for publication]. *IEEE Transactions on Education*.

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach.* Upper Saddle River, NJ: Pearson Education.

Tambe, M. (1998). Implementing agent teams in dynamic multiagent environments. *Applied Artificial Intelligence, 12*(2-3), 189-210.

KEY TERMS

Agent Framework: A software agent framework is a program or code library that provides a comprehensive set of capabilities that are used to develop and support software agents.

Autonomous Software Agent: An agent with the ability to anticipate changes in the environment so that the agent will change its behavior to improve the chance that it can continue performing its intended function.

Distributed Computing: The process of using a number of separate but networked computers to solve a single problem.

Emergent Behavior: The behavior that results from the interaction between a multitude of entities, where the observed behavior is not present in any single entity in the multitude comprising the system that shows emergent behavior.

Heterogeneous Agents: Agents of a multi-agent sSystem that differ in the resources available to them in the problem-solving methods and expertise they use, or in everything except in the interaction language they use.

Homogeneous Agents: Agents of a multi-agent system that are designed in an identical way and have a priori of the same capabilities.

Multi-Agent System (MAS): A multi-agent system is a collection of software agents that interact. This interaction can come in any form, including competition. The collection's individual entities and the interaction behavior together comprise the multi-agent system.

Software Agent: A self-contained piece of software that runs on an agent framework with an intended function to accomplish a simple goal.

Application of Genetic Algorithms for QoS Routing in Broadband Networks

Leonard Barolli
Fukuoka Institute of Technology, Japan

Akio Koyama
Yamagata University, Japan

INTRODUCTION

The networks of today are passing through a rapid evolution and are opening a new era of Information Technology (IT). In this information age, customers are requesting an ever-increasing number of new services, and each service will generate other requirements. This large span of requirements introduces the need for flexible networks. Also, future networks are expected to support a wide range of multimedia applications which raises new challenges for the next generation broadband networks. One of the key issues is the Quality of Service (QoS) routing (Baransel, Dobosiewicz, & Gburzynski, 1995; Black, 2000; Chen & Nahrstedt, 1998; Wang, 2001). To cope with multimedia transmission, the routing algorithms must be adaptive, flexible, and intelligent (Barolli, Koyama, Yamada, & Yokoyama, 2000, 2001). Use of intelligent algorithms based on Genetic Algorithm (GA), Fuzzy Logic (FL), and Neural Networks (NN) can prove to be efficient for telecommunication networks (Douligeris, Pistillides, & Panno, 2002). As opposed to non-linear programming, GA, FL and NN use heuristic rules to find an optimal solution.

In Munemoto, Takai, and Sato,(1998), a Genetic Load Balancing Routing (GLBR) algorithm is proposed and its behavior is compared with conventional Shortest Path First (SPF) and Routing Information Protocol (RIP). The performance evaluation shows that GLBR has a better behavior than SPF and RIP. However, in Barolli, Koyama, Motegi, and Yokoyama (1999), we found that GLBR genetic operations are complicated. For this reason, we proposed a new GA-based algorithm called Adaptive Routing method based on GA (ARGA). ARGA has a faster routing decision than GLBR. But, the ARGA and GLBR use only the delay time as a parameter for routing.

In order to support multimedia communication, it is necessary to develop routing algorithms which use for routing more than one QoS metric such as throughput, delay, and loss probability (Barolli, Koyama, Suganuma, & Shiratori, 2003; Barolli, Koyama, Sawada, Suganuma, & Shiratori, 2002b; Matsumoto, Koyama, Barolli, & Cheng, 2001). However, the problem of QoS routing is difficult, because the distributed applications have very diverse QoS constraints on delay, loss ratio, and bandwidth. Also, multiple constraints make the routing problem intractable and finding a feasible route with two independent path constraints is NP-complete (Chen & Nahrstedt, 1998). In this article, we propose two GA-based routing algorithms for multimedia communication: the first one called ARGAQ uses two QoS parameters mixed into a single measure by defining a function; and the second one is based on multi-purpose optimization and is used for multiple metrics QoS routing.

USE OF GA FOR NETWORK ROUTING

The GA cycle is shown in Figure 1. First, an initial population is created as a starting point for the search. Then, the fitness of each individual is evaluated with respect to the constraints imposed by the problem. Based on each individual's fitness, a selection mechanism chooses "parents" for the crossover and mutation. The crossover operator takes two chromosomes and swaps part of their genetic information to produce new chromosomes. The mutation operator introduces new genetic structures in the

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

population by randomly modifying some of genes, helping the algorithm to escape from local optimum. The offspring produced by the genetic manipulation process are the next population to be evaluated. The creation-evaluation-selection-manipulation cycle repeats until a satisfactory solution to the problem is found, or some other termination criteria are met (Gen, 2000; Goldberg, 1989). The main steps of GA are as follows.

1. Supply a population P0 of N individuals (routes) and respective function values;
2. i ← 1;
3. P'i ← selection_function (Pi-1);
4. Pi ← reproduction_function (P'i);
5. Evaluate (Pi);
6. i ← i+1;
7. Repeat step 3 until termination;
8. Print out the best solution (route).

The most important factor to achieve efficient genetic operations is gene coding. In the case when GA is used for routing and the algorithm is a source-based algorithm, a node which wants to transmit the information to a destination node becomes the source node. There are different coding methods of network nodes as GA genes. A simple coding method is to map each network node to a GA gene. Another one is to transform the network in a tree network with the source node as the root of tree. After that, the tree network may be reduced in the parts where are the same routes. Then, in the reduced tree network, the tree junctions may be coded as genes.

After the crossover and mutation, the elitist model is used. Based on the elitist model, the route which has the highest fitness value in a population is left intact in the next generation. Therefore, the best value is always kept and the routing algorithm can converge very fast to the desired value. The offsprings produced by the genetic operations are the next population to be evaluated. The genetic operations are repeated until the initialized generation size is achieved or a route with a required optimal value is found.

OUTLINE OF PREVIOUS WORK

In this section, we will explain ARGA and GLBR algorithms. In the GLBR, the genes are put in a chromosome in the same order the nodes form the communication route, so the chromosomes have different size. If genetic operations are chosen randomly, a route between two adjacent nodes may not exist and some complicated genetic operations should be carried out to find a new route. Also, because the individuals have different size, the crossover operations become complicated. On the other hand, in ARGA the network is expressed by a tree network and the genes are expressed by tree junctions. Thus, the routing loops can be avoided. Also, the length of each chromosome is the same and the searched routes always exist. Therefore, there is no need to check their validity (Barolli, Koyama, Yamada, Yokoyama, Suganuma, & Shiratori, 2002a). To explain this procedure, we use a small network with 8 nodes as shown in Figure 2. Node A is the source node and node H is the destination node. All routes are expressed by the network tree model shown in Figure 3. The shaded areas show the same routes from node C to H. Therefore, the network tree model of Figure 3 can be reduced as shown in Figure

Figure 1. GA cycle

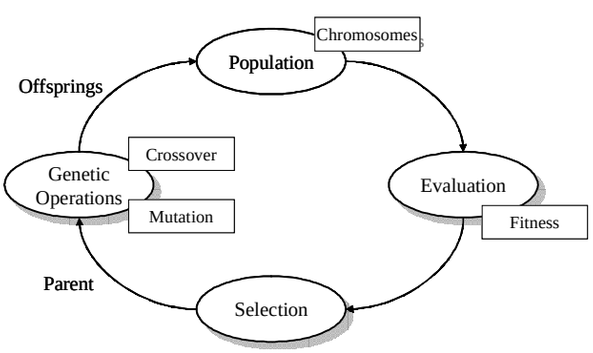


Figure 2. Network example with 8 nodes

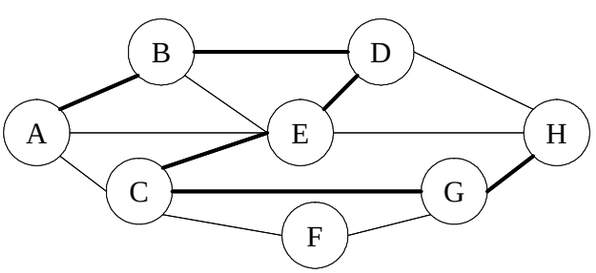


Figure 3. Tree network

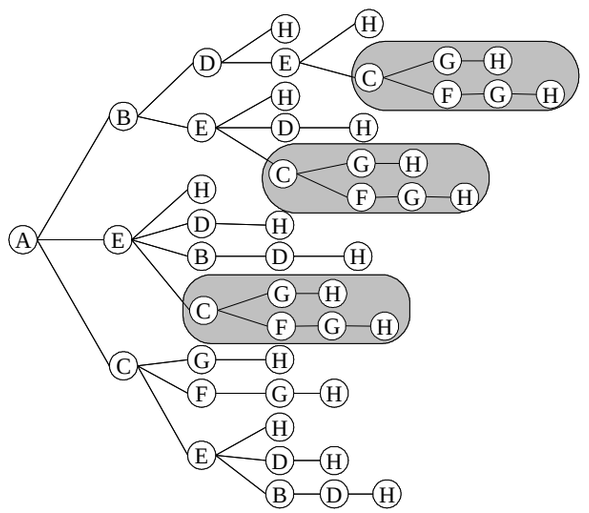


Figure 4. Reduced tree network

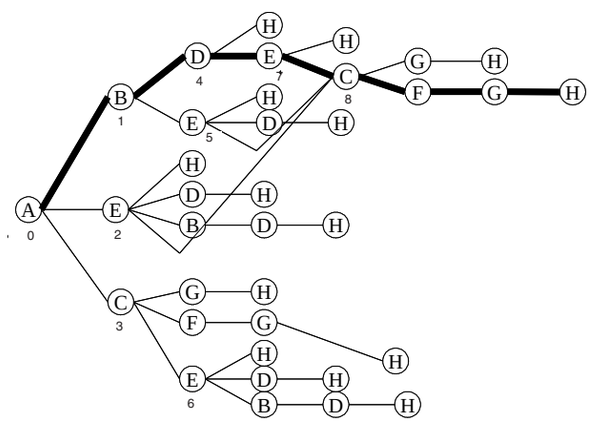
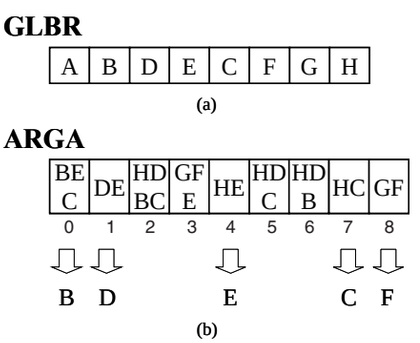


Figure 5. GLBR and ARGA gene coding



4. In this model, each tree junction is considered as a gene and the route is represented by the chromosome. Figure 5(a) and Figure 5(b) show the route B-D-E-C-F-G-H, for GLBR and ARGA, respectively.

OUTLINE OF QoS ROUTING ALGORITHMS

The routing algorithms can be classified into: single metric, single mixed metric, and multiple metrics. In following, we will propose a single mixed (ARGAQ) and a multiple metrics GA-based QoS routing algorithms

ARGAQ Algorithm

In ARGA and GLBR algorithms, the best route was decided considering only the delay time. The ARGAQ is a unicast source-based routing algorithm and uses for routing two parameters: the Delay Time (DT) and Transmission Success Rate (TSR). Let consider a network as shown in Figure 6. The node A is a source node and node B is the destination node. Let node A sends 10 packets to node B. The total TSR value for Figures 6(a) and Figure 6(b) is calculated by Eq.(1) and Eq.(2), respectively.

$$10 \times 0.9 \times 0.9 \times 0.9 \times 0.9 = 6.561 \quad (1)$$

$$10 \times 1.0 \times 1.0 \times 0.6 \times 1.0 = 6.000 \quad (2)$$

The best route in this case is that of Figure 6(a), because the total TSR is higher compared with that of Figure 6(b).

Let consider another example, when the values of DT and TSR are considered as shown in Figure 7. The value of T parameter is decided as follows.

$$T = \sum DTi \,/\, \prod TSRi \quad (3)$$

where "i" is link number which varies from 1 to n.

When node A wants to communicate with node D, there are two possible routes: "A-B-D" and "A-C-D". The T value for these routes are calculated by Eq.(4) and Eq.(5), respectively.

Figure 6. An example of TSR calculation

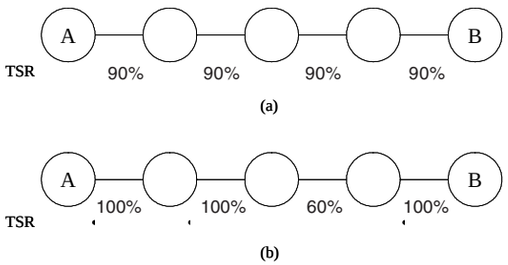
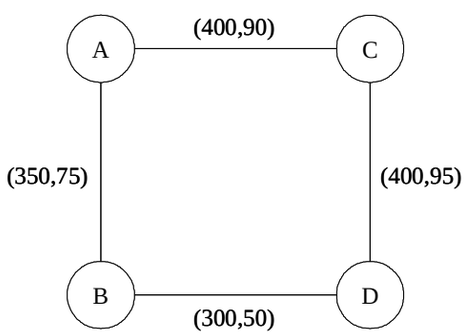


Figure 7. A network example for QoS routing



$$TA\text{-}B\text{-}D = (350 + 300)/(75 \times 50) = 650/3750 = 0.1733 \quad (4)$$

$$TA\text{-}C\text{-}D = (400 + 400)/(90 \times 95) = 800/8550 = 0.0468 \quad (5)$$

The delay time of "A-B-D" route is lower than "A-C-D" route, but the T value of "A-C-D" route is lower than "A-B-D", so "A-C-D" route is the better one. This shows that a different candidate route can be found when two QoS parameters are used for routing.

Multi-Purpose Optimization Method

The proposed method uses the multi-division group model for multi-purpose optimization. The global domain is divided in different domains and each GA individual evolves in its domain as shown in Figure 8. Figure 9 shows an example of Delay Time (DT) and Communication Cost (CC). The shaded area is called "pareto solution". The individuals near pareto solution can be found by exchange the solutions of different domains.

The structure of proposed Routing Search Engine (RSE) is shown in Figure 10. It includes two search engines: Cache Search Engine (CSE) and Tree Search Engine (TSE). Both engines operate independently, but they cooperate together to update the route information. When the RSE receives a request, it forwards the request to CSE and TSE. Then, the CSE and TSE search in parallel to find a route satisfying the required QoS. The CSE searches for a route in the cache database. If it finds a QoS route sends it to RSE. If a QoS route isn't found by CSE, the route found by TSE is sent to RSE. The CSE is faster than TSE, because the TSE searches for all routes in its domain using a GA-based routing. The database should be updated because the network traffic and the network state change dynamically. The database update is shown in Figure 11. After CSE finds a route in the database, it checks whether this route satisfies or not the required QoS. If the QoS is not satisfied, then this route is deleted from the database. Otherwise, the route is given higher priority and can be searched very fast during the next search.

SIMULATION RESULTS

Matsumoto et al. (1998) show the performance evaluation of GLBR, SPF and RIP. In this article, we evaluate by simulations the performance of the GA-based routing algorithms.

ARGAQ Simulation Results

We carried out many simulations for different kinds of networks with different number of nodes, routes and branches as shown in Table 1. We implemented a new routing algorithm based on GLBR and called it GLBRQ. Then, we compare the results of ARGAQ and GLBRQ.

First, we set in a random way the DT and TSR in each link of the network. Next, we calculate the value T, which is the ratio of DT with TSR. This value is used to measure the individual fitness. The genetic operations are repeated until a route with a small T value is found or the initialized generation

Figure 8. Multiple-purpose optimization

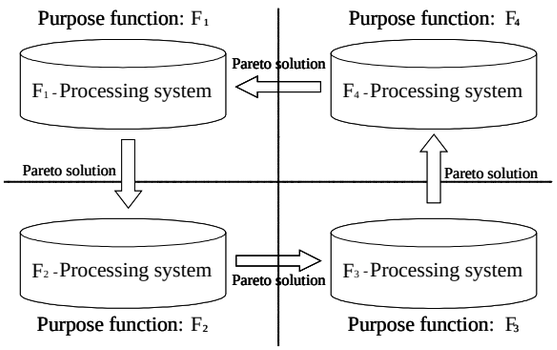


Figure 9. Pareto solution for DT and CC

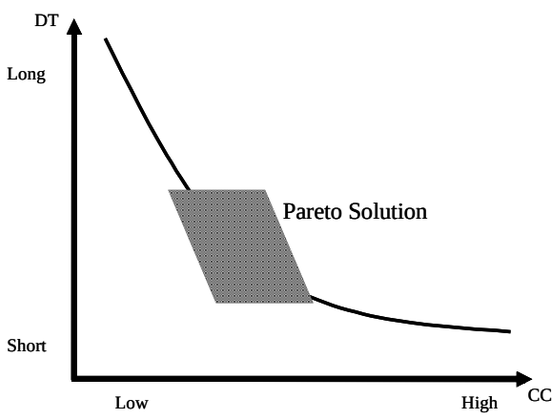
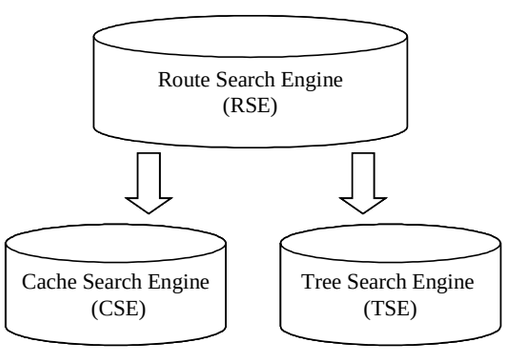


Figure 10. RSE structure



size is achieved. For the sake of comparison, we use the same parameters and the population size. Performance behavior of ARGAQ and GLBRQ is shown in Figure 12. The rank is decided based on the value of fitness function T. When the rank is low the fitness value is low. This means, the selected route has a low delay and a high transmission rate. The average rank value of ARGAQ is lower than average rank value of GLBRQ for the same generation number. This means GLBRQ needs more genetic operations to find a feasible route. Therefore, the search efficiency of ARGAQ is better than GLBRQ.

In Table 2, Rank is the average rank to find a new route; Gen is the average number of generations to find a new route; Fail is the rate that a new route was not found (percent); and Ref is the average number of individuals refereed in one simulation. Considering the results in Table 2, the ARGAQ can find a new route by using few generations than GLBRQ. For the network with 30 nodes, the failure for GLBRQ was about 14 percent. For the network with 30 nodes the failure rate is about two times more than the network with 35 nodes. This shows that by increasing the network scale the ARGAQ shows better behavior than GLBRQ.

In Table 3, we show the simulation results of ARGAQ and ARGA. The TA means the average rank value of T parameter, DA means the average rank value of delay, TSRA means the average rank value of TSR parameter, GSA means the average value of generation number, and GOTA means the average value of genetic processing time. The genetic operations processing time of ARGA is better than ARGAQ. However, the difference is very small (see parameter GOTA). In the case of ARGAQ, both DA and TSRA values are optimized. However, in the case of ARGA only one QoS parameter is used. Thus, only DA value is optimized, the TSRA value is large. Therefore, the selected route is better from the QoS point of view when two QoS parameters are used.

TSE Simulation Results

For the TSE simulation, we use a network with 20 nodes as shown in Figure 13. First, we set in a random way the DT and CC in each link of network. The RSE generates in random way the values of the

Figure 11. Cache database update

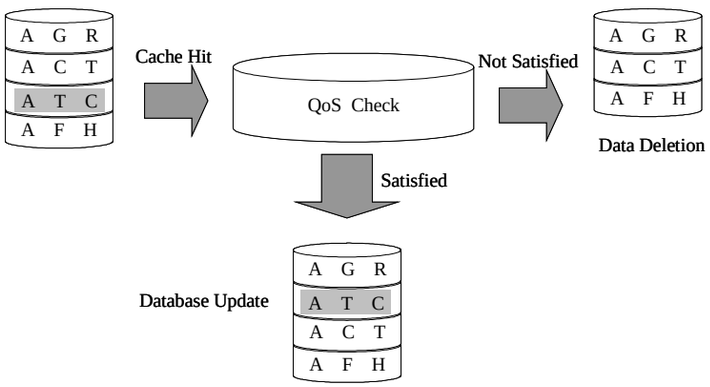


Table 1. Number of nodes, routes, and branches

Nodes	20	30	35
Routes	725	11375	23076
Branches	33	85	246

Figure 12. Performance of ARGAQ and GLBRQ

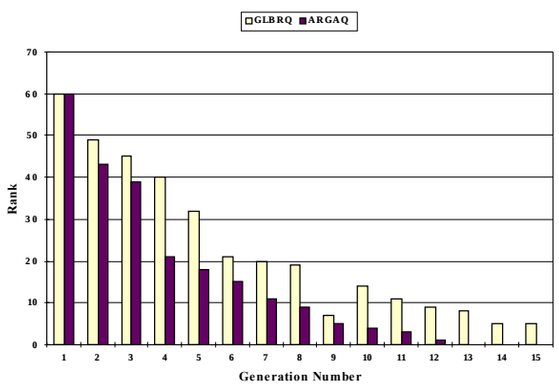


Table 2. Performance for different parameters

Nodes	Method	Rank	Gen	Fail	Ref
20	GLBRQ	5.5	33.5	6	54.32
	ARGAQ	5.62	8	0	26.3
30	GLBRQ	8.8	69.94	14	123.12
	ARGAQ	6.44	53.3	8	100.94
35	GLBRQ	6.12	55.52	6	103.84
	ARGAQ	5.38	28.72	0	65.62

Table 3. Comparison between ARGAQ and ARGA

Method	TA	DA	TSRA	GSA	GOTA
ARGAQ	4.47	10.52	9.36	9	85.78
ARGA	-	4.66	70.6	8.33	69.04

required QoS and the destination node. Next, the CSE and TSE search in parallel to find a route. If the CSE finds a route in the cache database, it checks whether it satisfies the QoS or not. If so, this route is sent back to the RSE. Otherwise, the route is put as a new individual in the gene pool. If CSE doesn't find a QoS route, the route found by TSE is sent to RSE. The genetic operations are repeated until a solution is found or the number of 200 generations is achieved. In Table 4 we show the TSE simulation results. If there are few individuals in the population, the GN which shows the number of generations needed to find a solution becomes large. When the number of individuals is high, the GN to find a solution becomes small. However, when the number of individuals is 12 and 16, the difference is very small because some individuals become the same in the gene pool. Also, when the exchange interval is short the solution can be found very fast. This shows that by exchanging the individuals the algorithm can approach very quickly to the pareto solution.

Comparison between GA-Based Routing Algorithms

Table 5 shows the comparison between GA-based routing algorithms. The GLBR and ARGA use as the Routing Parameter (RP) DT, ARGAQ uses DT and TSR, and TSE uses DT and CC. The GLBR uses for Gene Coding (GC) the nodes of network, while ARGA, ARGAQ and TSE use the tree junctions. By using the network nodes as gene, the GLBR may enter in routing loops. Also, the searched route may not exist, so the algorithm after searching a route

Figure 13. Network model with 20 nodes

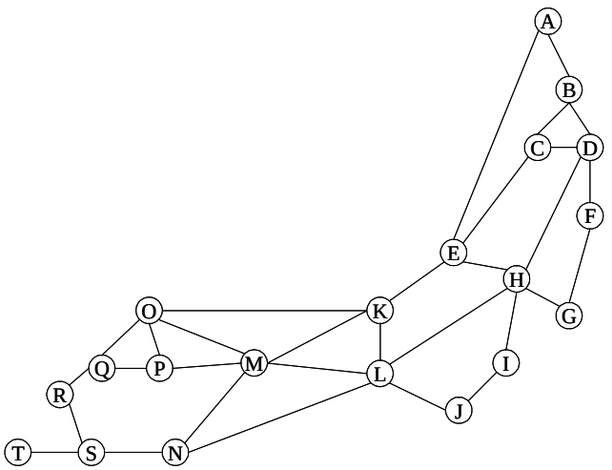


Table 4. Time needed for one generation (ms)

Number of Individuals	GN Exchange			
	3	5	7	10
4	44.43	50.45	46.19	55.59
8	26.83	28.01	40.26	31.17
12	23.55	26.49	26.04	26.71
16	22.22	22.23	23.25	24.04

Table 5. GA-based routing algorithms comparision

Method	RP	GC	RS	AC	RSCM
GLBR	DT	Network Nodes	Source	Low	Single Metric
ARGA	DT	Tree Junctions	Source	Low	Single Metric
ARGAQ	DT, TSR	Tree Junctions	Source	Middle	Single Mixed Metric
TSE	DT, CC	Tree Junctions	Source	Middle	Multiple Metrics

should check whether the route exists or not. If the searched route does not exist, the GLBR should search for another route. Thus, the searching time increases. Three other algorithms by using as gene the tree junction can avoid the routing loops and always the route exist. So there is not need to check the route existence. All four algorithms use as Routing Strategy (RS) the source routing thus they are considered source-based routing methods. Considering the algorithm complexity, the GLBR and ARGA have a low complexity, because they use only one parameter for routing. The complexity of ARGAQ and TSE is higher than GLBR and ARGA. The last comparison is about the Routing Selection Criterion Metrics (RSCM). The GLBR and ARGA use single metric (DT). Thus, they can not be used for QoS routing. The ARGAQ uses a single mixed metric (T), which is the ratio of DT and TSR. By using the single mixed metric, the ARGAQ can be used only as an indicator because it does not contain sufficient information to decide whether user QoS requirements can be met or not. Another problem with ARGAQ has to do with mixing of parameters of different composition rules, because may be not simple composition rule at all. The TSE uses multiple metrics for route selection. In the proposed method, the DT and CC have trade-off relation and to get the composition rule the TSE uses pareto solution method. In this paper, we used only two parameters for QoS routing. However, the TSE different from ARGAQ can use for routing multiple QoS metrics.

We intend to use the proposed algorithms for small-scale networks. For large-scale networks, we have implemented a distributed routing architecture based on cooperative agents (Barolli et al., 2002a). In this architecture, the proposed algorithms will be used for intra-domain routing.

CONCLUSION

We proposed two GA-based QoS routing algorithms for multimedia applications in broadband networks. The performance evaluation via simulations shows

that ARGAQ has a faster response time and simple genetic operations compared with GLBRQ. Furthermore, ARGAQ can find better QoS routes than ARGA. The evaluation of the proposed multi-purpose optimization method shows that when there are few individuals in a population, the GN becomes large. When the exchange interval of individuals is short, the solution can be found very fast and the algorithm can approach very quickly to the pareto solution. The multi-purpose optimization method can support QoS routing with multiple metrics. In this article, we carried out the simulations only for two QoS metrics. In the future, we would like to extend our study to use more QoS metrics for routing.

REFERENCES

Baransel, C. Dobosiewicz, W., & Gburzynski, P. (1995). Routing in multihop packet switching networks: GB/s challenge. *IEEE Network, 9*(3), 38-60.

Barolli, L., Koyama, A., Motegi, S., & Yokoyama, S. (1999). Performance evaluation of a genetic algorithm based routing method for high-speed networks. *Trans. of IEE Japan*, *119-C*(5), 624-631.

Barolli, L., Koyama, A., Sawada, H.S., Suganuma, T., & Shiratori, N. (2002b). A new QoS routing approach for multimedia applications based on genetic algorithms. *Proceedings of CW2002,* Tokyo, Japan, (pp. 289-295).

Barolli, L., Koyama, A., Suganuma, T., & Shiratori, N. (2003). A genetic algorithm based QoS routing method for multimedia communications over high-speed networks. *IPSJ Journal, 44*(2), 544-552.

Barolli, L., Koyama, A., Yamada, T., & Yokoyama, S. (2000). An intelligent policing-routing mechanism based on fuzzy logic and genetic algorithms and its performance evaluation. *IPSJ Journal, 41*(11), 3046-3059.

Barolli, L., Koyama, A., Yamada, T., & Yokoyama, S. (2001). An integrated CAC and routing strategy for high-speed large-scale networks using cooperative agents. *IPSJ Journal, 42*(2), 222-233.

Barolli, L., Koyama, A., Yamada, T., Yokoyama, S., Suganuma, T., & Shiratori, N. (2002a). An intelligent routing and CAC framework for large scale networks based on cooperative agents. *Computer Communications Journal, 25*(16), 1429-1442.

Black, U. (2000). *QoS in wide area networks.* Prentice Hall PTR.

Chen, S. & Nahrstedt, K. (1998). An overview of quality of service routing for next-generation high-speed networks: Problems and solutions. *IEEE Network, Special Issue on Transmission and Distribution of Digital Video, 12*(6), 64-79.

Douligeris, C., Pistillides, A., & Panno, D. (Guest Editors) (2002). Special issue on computational intelligence in telecommunication networks. *Computer Communications Journal, 25*(16).

Gen, M. & Cheng, R. (2000). *Genetic algorithms & engineering optimization*. John Wiley & Sons.

Goldberg, D.E. (1989). *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley.

Matsumoto, K., Koyama, A., Barolli, L., & Cheng, Z (2001). A QoS routing method for high-speed networks using genetic algorithm. *IPSJ Journal, 42*(12), 3121-3129.

Munemoto, M., Takai, Y., & Sato, Y. (1998). An adaptive routing algorithm with load balancing by a genetic algorithm. *Trans. of IPSJ, 39*(2), 219-227.

Wang, Z. (2001). *Internet QoS: Architectures and mechanisms for quality of service.* Academic Press.

KEY TERMS

Broadband Networks: Networks which operate at a wide band of frequencies. In these communications networks, the bandwidth can be divided and shared by multiple simultaneous signals (as for voice or data or video).

Genetic Algorithm: An evolutionary algorithm which generates each individual from some encoded form known as a “chromosome” or “genome”. Chromosomes are combined or mutated to breed new individuals.

Heurictic Rule: A commonsense rule (or set of rules) intended to increase the probability of solving some problems.

Intelligent Algorithms: Human-centered algorithms, which have the capacity for thought and reason especially to a high degree.

Multimedia Transmission: Transmission that combines media of communication (text, graphics, sound, etc.)

QoS: The ability of a service provider (network operator) to support the application requirements with regard to four services categories: bandwidth, delay, jitter, and traffic loss.

Unicast: One to one communication.

Application Service Providers

Vincenzo Morabito
Bocconi University, Italy

Bernardino Provera
Bocconi University, Italy

INTRODUCTION

Until recently, the development of information systems has been ruled by the traditional "make or buy" paradigm (Williamson, 1975). In other words, firms could choose whether to develop particular applications within their organizational structure or to acquire infrastructures and competences from specialized operators. Nevertheless, the Internet's thorough diffusion has extended the opportunities that firms can rely upon, making it possible to develop a "make, buy, or rent" paradigm. Application service providers represent the agents enabling this radical change in the IS scenario, providing clients with the possibility to rent specifically-tailored applications (Morabito, 2001; Pasini, 2002).

Our research aims at analyzing ASPs in terms of organizational characteristics, value chain, and services offered. Moreover, we analyze the set of advantages that ASPs can offer with respect to cost reductions, technical efficiency, implementation requirements, and scalability. Finally, we describe the major challenges these operators are currently facing and how they manage to overcome them.

BACKGROUND

ASPs are specialized operators that offer a bundle of customized software applications from a remote position through the Internet, in exchange for a periodic fee. ASPs provide for the maintenance of the system network and for upgrading its offer on a continuous basis. The historical development of ASPs follows the diffusion of the Internet. Early actors began to operate around 1998 in the U.S., while a clear definition of their business model has only recently come to shape. As opposed to traditional outsourcing, the ASP offer is based on a one-to-many relationship that allows different clients to gain access to a defined set of applications through a browser interface (Factor, 2001).

MAIN FOCUS

Information and Communication Technology (ICT) is widely believed to represent a crucial determinant of an organization's competitive positioning and development (Brown & Hagel, 2003; Varian, 2003). On the other hand, companies often face the problem of aligning corporate strategies with ICT resources and capabilities (Carr, 2003), in order to rely on the necessary applications at the right time and place, allowing for the effective implementation of business strategies. The inability to match corporate strategy and ICT capabilities might lead to efficiency and efficacy losses. In particular, Information Systems are among the organizational functions most affected by the organizational and strategic changes induced by the Internet. Historically, firms could rely on two possibilities for designing and implementing Information Systems. The first option is to develop applications internally with proprietary resources and competences. The second possibility is to acquire such solutions from specialized market operators. Despite the conceptual relevance of this distinction, the range of applications currently available on the market is ample and encompasses a series of hybrid solutions that lie on a continuum between the make and the buy option (Afuah, 2003; Bradach & Eccles, 1989; Hennart, 1993; Poppo & Zenger, 1998). In that sense, standard outsourcing relations hardly ever take the shape of pure spot solutions. On the contrary, outsourcing contracts often develop into long-run partnerships (Willcocks & Lacity, 1999). Therefore, the ASP model can be conceived as a hybrid solution located on the continuum between market and hierarchy (Williamson, 1975). Nevertheless, as shown in the following paragraphs, the ASP option presents

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

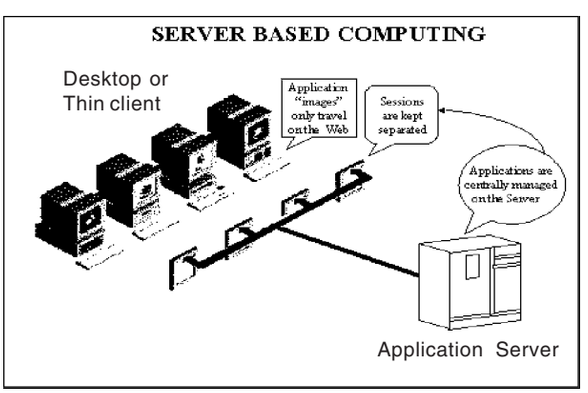
particular, stand-alone peculiarities and features such as to make it different from traditional make or buy models and to acquire a level of conceptual legitimacy in itself.

The ASP model is based on two key technologies: Internet and server-based computing. The first technology represents the building network of the system, while server-based computing allows many remote clients to obtain access to applications running on a single server. The functioning mechanism is quite simple: the server manages as many protected and separate sessions as the number of logged-in users. Only the images of the interface, client-inserted data and software upgrades "travel" on the Internet, while all applications reside on the server, where all computations also take place. Figure 1 provides a visual representation of the functioning of a server based computing system.

Client firms can rent all kinds of business applications, ranging from very simple to highly complex ones, as described below:

- Personal applications, allowing individual analysis of basic, everyday activities and tasks (e.g., Microsoft Office).
- Collaborative applications, supporting the creation of virtual communities (e.g., groupware, e-mail, and video-conference systems).
- Commercial applications, aimed at creating and maintaining e-commerce solutions.
- Customer Relationship Management systems (e.g., customer service, sales force automation, and marketing applications).

Figure 1. Server-based computing technology (Source: our elaboration)



- Enterprise Resource Planning, applications aimed at the automation of all business processes, at any organizational level (e.g., infrastructure management, accounting, human resources management, and materials planning).
- Analytical applications that allow for the analysis of business issues (risk analysis, financial analysis, and customer analysis).

Along with these applications, ASPs offer a wide array of services, as reported below:

- Implementations services that are required in order to align applications and business processes. These services include, for example, data migration from previous systems to the ASP server and employees' training.
- Data centre management services, aimed at assuring the reliability and security of hardware and software infrastructure, as well as transferred data.
- Support services, delivered on a non-stop basis, in order to solve technical and utilization problems.
- Upgrading services, aimed at aligning applications with evolving business needs and environmental change.

ASPs can hardly be fit into a single, monolithic categorization (Seltsikas & Currie, 2002). In fact, operators can be grouped into different classes, according to their offer system and market origins (Chamberlin, 2003). The first category includes enterprise application outsourcers, which are traditional operators in the field of IT outsourcing that deliver ASP services. They can rely on profound process knowledge, sound financial resources and wide geographic coverage. On the other hand, their great size can have negative impacts on deployment time, overall flexibility, and client management.

The second category of actors refers to pure-play ASPs, that usually demonstrate the highest degree of technical efficiency and competency in application infrastructure design and implementation. As opposed to enterprise application outsourcers, they are flexible, fast at deploying solutions and extremely attentive towards technology evolution, although they might be hampered by financial constraints and limited visibility.

The third class of operators includes independent software vendors, which can decide to license their products in ASP modality. These firms are extremely competent, technically skilful, and financially stable. On the other hand, they often lack experience in supporting clients in a service model and can be really competitive only when offering their own specialized sets of applications.

The final category of actors refers to Net-Native ASPs, smaller operators extremely agile and flexible, offering standard repeatable solutions. On the other hand, ASPs offer point solutions, are often financially restrained, partially visible and unable to customize their offer.

In order to ensure adequate service levels, ASPs must interact with a complex network of suppliers, that include hardware and software producers (or independent software vendors), technology consultants and connectivity suppliers. Software vendors generally offer ASPs particular licensing conditions, such as fees proportional to the number of users accessing the applications. Moreover, in order not to lose contact with previous clients, many software producers engage in long-term business partnerships with ASPs. Hardware vendors often develop strategic relationships with ASPs too, as the latter are interested in buying powerful servers, with advanced data storage and processing capabilities. Technology consultants are important actors as they can include ASPs' solutions in their operating schemes. Connectivity suppliers as Network Service Providers can decide whether to team up with an ASP or to offer themselves ASP solutions. In conclusion, ASPs rely on a distinct business model, which can be defined as "application renting", where the ability to coordinate a complex network of relationships is crucial.

The ASP business model (or "rent" option) is different from that of traditional outsourcing (or "buy" option) due to three main reasons (Susarla, Barua & Whinston, 2003). First of all, an ASP centrally manages applications, in its own data centre, rather than at the clients' location. Second, ASPs offer a one-to-many service, given that the same bundle of applications is simultaneously accessible to several users. On the contrary, traditional outsourcing contracts are a one-to-one relationship in which the offer is specifically tailored to suit clients' needs. The third main difference, in fact, is that ASPs generally offer standard products, although they might include application services specifically conceived for particular client categories.

Adopting the "rent" option allows firms to benefit from a wide set of advantages. First of all, the ASP model can remarkably reduce the operating costs of acquiring and managing software, hardware and relative know-how. In particular, the total cost of IT ownership notably decreases (Seltsikas & Currie, 2002). Second, costs become more predictable and stable, as customers are generally required to pay a monthly fee. These two advantages allow for the saving of financial resources that can be profitably reinvested in the firm's core business. The third benefit refers to the increase in technical efficiency that can be achieved by relying on a specialized, fully competent operator. Moreover, with respect to developing applications internally or buying tailored applications from traditional outsourcers, implementation time considerably decreases, allowing firms to operate immediately. Finally, ASPs offer scalable applications that can be easily adjusted according to the clients' evolving needs.

In conclusion, the ASP model leads to minimize complexity, costs, and risks, allowing also small firms to gain access to highly evolved business applications (as, for example, ERP systems) rapidly and at reasonable costs. Nevertheless, the adoption of an ASP system might involve potential risks and resistances that must be attentively taken into account (Kern, Willcocks & Lacity, 2002). We hereby present the most relevant issues, as well as explain how ASPs try to overcome them (Soliman, Chen & Frolick, 2003).

Clients are often worried about the security of information exchanged via the Web, with special reference to data loss, virus diffusion, and external intrusions. Operators usually respond by relying on firewalls and virtual private networks, with ciphered data transmission. Another key issue refers to the stability of the Internet connection, which must avoid sudden decreases in download time and transmission rates. In order to ensure stable operations, ASPs usually engage in strategic partnerships with reliable carriers and Internet Service Providers. Moreover, clients often lament the absence of precise agreements on the level of service that operators guarantee (Pring, 2003). The lack of clear contractual commitment might seriously restrain potentially interested clients from adhering to the ASP model.

Therefore, many operators include precise service level agreements clauses in order to reassure clients about the continuity, reliability, and scalability of their offer. Finally, the adoption of innovative systems architecture on the Internet might be hampered by cultural resistances, especially within smaller firms operating in traditional, non technology-intensive environments. In this case, ASPs offer on the spot demonstrations, continuous help desk support, attentive training and simplified application bundles not requiring complex interaction processes.

FUTURE TRENDS

Regarding future development, some observers have predicted that, by 2004, 70 percent of the most important firms in business will rely on some sort of outsourcing for their software applications (Chamberlin, 2003). The choice will be between traditional outsourcers, niche operators, offshore solutions, and ASPs. Moreover, according to a research carried out by IDC and relative to the United States alone, the ASP market is to grow from $1.17 billion in 2002 to $3.45 billion by 2007. Similar growth trends are believed to apply to Europe as well (Lavery, 2001; Morganti, 2003). Other observers believe that the ASP market will be affected by a steady process of concentration that will reduce competing firms from over 700 in 2000 to no more than 20 in the long run (Pring, 2003).

CONCLUSION

In conclusion, we argue that ASPs represent a new business model, which can be helpful in aligning corporate and IT strategies. The "rent" option, in fact, involves considerable advantages in terms of cost savings and opportunities to reinvest resources and attention in the firm's core activities. As many other operators born following the rapid diffusion of the internet, ASPs also experienced the wave of excessive enthusiasm and the dramatic fall of expectations that followed the burst of the Internet bubble. Nonetheless, as opposed to other actors driven out of business due to the inconsistency of their business models, ASPs have embarked upon a path of moderate yet continuous growth. Some observers believe that ASPs will have to shift their focus from delivery and implementation of software applications to a strategy of integration with other key players as, for example, independent server vendors (Seltsikas & Currie, 2002). As the industry matures, reducing total costs of ownership might simply become a necessary condition for surviving in the business, rather than an element of competitive advantage. ASPs should respond by providing strategic benefits as more secure data, better communications, attractive service-level agreements and, most important, integration of different systems. The ASP business model might involve a strategy of market segmentation, including customized applications from more than one independent software vendor, in order to offer solutions integrating across applications and business processes.

REFERENCES

Afuah, A. (2003). Redefining firms boundaries in the face of the Internet: Are firms really shrinking? *Academy of Management Review*, *28*(1), 34-53.

Bradach, J. & Eccles R. (1989). Price, authority, and trust: From ideal types to plural forms. *Annual Review of Sociology*, *15*, 97-118.

Brown, J.S. & Hagel III, J. (2003). Does IT matter? *Harvard Business Review*, July.

Carr, N.G. (2003). IT doesn't matter. *Harvard Business Review*, May.

Chamberlin, T. (2003). *Management update: What you should know about the Application Service Provider Market.* Market Analysis, Gartner Dataquest.

Factor, A. (2001). *Analyzing application service provider*s. London: Prentice Hall.

Hennart, J.F. (1993). Explaining the swollen middle: Why most transactions are a mix of "market" and "hierarchy". *Organizations Science*, *4*, 529-547.

Kern, T., Willcocks, L.P., & Lacity M.C. (2002). Application service provision: Risk assessment and mitigation. *MIS Quarterly Executive*, *1*, 113-126.

Lavery, R. (2001). The ABCs of ASPs. *Strategic Finance*, *52*, 47-51.

Morabito, V. (2001). I sistemi informativi aziendali nell'era di Internet: gli Application Service Provider, in Demattè (a cura di), *E-business: Condizioni e strumenti per le imprese che cambiano*, ETAS, Milano.

Morganti, F. (2003). Parola d'ordine: pervasività, *Il Sole 24Ore*, 5/6/2003.

Pasini, P. (2002). *I servizi di ICT. Nuovi modelli di offerta e le scelte di Make or Buy*. Milano, Egea.

Poppo, L. & Zenger, T. (1998). Testing alternative theories of the firms: Transaction costs, knowledge-based and measurement explanation for make-or-buy decision in information services. *Strategic Management Journal,* 853-877.

Pring, B. (2003a). *2003 ASP Hype: Hype? What Hype?* Market Analysis, Gartner Dataquest.

Pring, B. (2003b). *The New ASO Market: Beyond the First Wave of M&A.* Market Analysis, Gartner Dataquest.

Seltsikas, P. & Currie, W. (2002). Evaluating the Application Service Provider (ASP) business model: The challenge of integration, *Proceedings of the 35th Hawaii International Conference on System Sciences*.

Soliman, K.S., Chen, L., & Frolick, M.N. (2003). ASP: Do they work? *Information Systems Management,* 50-57.

Susarla, A., Barua, A., & Whinston, A. (2003). Understanding the service component of application service provision: An empirical analysis of satisfaction with ASP services. *MIS Quarterly*, 27(1).

Varian, H. (2003). Does IT matter? *Harvard Business Review*, July.

Willcocks, L. & Lacity, M. (1999). *Strategic outsourcing of information systems: Perspectives and practices.* New York: Wiley & Sons.

Williamson, O.E. (1975). *Markets and hierarchies: Analysis and antitrust implications.* New York: Free Press.

Young, A. (2003). *Differentiating ASPs and traditional application outsourcing.* Commentary, Gartner Dataquest.

KEY TERMS

Application Outsourcing: Multiyear contract or relationship involving the purchase of ongoing applications services from an external service provider that supplies the people, processes, tools and methodologies for managing, enhancing, maintaining and supporting both custom and packaged software applications, including network-delivered applications (Young, 2003).

ASP (Application Service Providers): Specialized operators that offer a bundle of customized software applications from a remote position through the Internet, in exchange for a periodic fee.

CRM (Customer Relationship Management): Methodologies, softwares, and capabilities that help an enterprise manage customer relationships in an organized way.

ERP (Enterprise Resource Planning): Set of activities supported by multi-module application software that helps a manufacturer or other business manage the important parts of its business, including product planning, parts purchasing, maintaining inventories, interacting with suppliers, providing customer service, and tracking orders.

Network Service Providers: A company that provides backbone services to an Internet service provider, the company that most Web users use for access to the Internet. Typically, an ISP connects, at a point called Internet Exchange, to a regional Internet Service Provider that in turn connects to a Network Service Provider backbone.

Server-Based Computing (or Thin-Client Technology): Evolution of client-server systems in which all applications and data are deployed, managed and supported on the server. All of the applications are executed at the server.

Service Level Agreement (SLA): Contract between a network service provider and a customer that specifies, usually in measurable terms, what services the network service provider will furnish.

Assessing Digital Video Data Similarity

Waleed E. Farag
Zagazig University, Egypt

INTRODUCTION

Multimedia applications are rapidly spread at an ever-increasing rate, introducing a number of challenging problems at the hands of the research community. The most significant and influential problem among them is the effective access to stored data. In spite of the popularity of the keyword-based search technique in alphanumeric databases, it is inadequate for use with multimedia data due to their unstructured nature. On the other hand, a number of content-based access techniques have been developed in the context of image and video indexing and retrieval (Deb, 2004). The basic idea of content-based retrieval is to access multimedia data by their contents, for example, using one of the visual content features.

Most of the proposed video-indexing and -retrieval prototypes have two major phases: the database-population and retrieval phases. In the former one, the video stream is partitioned into its constituent shots in a process known as shot-boundary detection (Farag & Abdel-Wahab, 2001, 2002b). This step is followed by a process of selecting representative frames to summarize video shots (Farag & Abdel-Wahab, 2002a). Then, a number of low-level features (color, texture, object motion, etc.) are extracted in order to use them as indices to shots. The database-population phase is performed as an off-line activity and it outputs a set of metadata with each element representing one of the clips in the video archive. In the retrieval phase, a query is presented to the system that in turns performs similarity-matching operations and returns similar data back to the user.

The basic objective of an automated video-retrieval system (described above) is to provide the user with easy-to-use and effective mechanisms to access the required information. For that reason, the success of a content-based video-access system is mainly measured by the effectiveness of its retrieval phase. The general query model adapted by almost all multimedia retrieval systems is the QBE (query by example; Yoshitaka & Ichikawa, 1999). In this model, the user submits a query in the form of an image or a video clip (in the case of a video-retrieval system) and asks the system to retrieve similar data. QBE is considered to be a promising technique since it provides the user with an intuitive way of query presentation. In addition, the form of expressing a query condition is close to that of the data to be evaluated.

Upon the reception of the submitted query, the retrieval stage analyzes it to extract a set of features, then performs the task of similarity matching. In the latter task, the query-extracted features are compared with the features stored into the metadata, then matches are sorted and displayed back to the user based on how close a hit is to the input query. A central issue here is the assessment of video data similarity. Appropriately answering the following questions has a crucial impact on the effectiveness and applicability of the retrieval system. How are the similarity-matching operations performed and on what criteria are they based? Do the employed similarity-matching models reflect the human perception of multimedia similarity? The main focus of this article is to shed the light on possible answers to the above questions.

BACKGROUND

An important lesson that has been learned through the last two decades from the increasing popularity of the Internet can be stated as follows: "[T]he usefulness of vast repositories of digital information is limited by the effectiveness of the access methods" (Brunelli, Mich, & Modena, 1999). The same analogy applies to video archives; thus, many researchers are starting to be aware of the significance of providing effective tools for accessing video databases. Moreover, some of them are proposing various techniques to improve the quality, effectiveness, and robustness of the retrieval system. In the following, a quick review of these techniques is introduced with emphasis on various approaches for evaluating video data similarity.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

One important aspect of multimedia-retrieval systems is the browsing capability, and in this context some researchers proposed the integration between the human and the computer to improve the performance of the retrieval stage. In Luo and Eleftheriadis (1999), a system is proposed that allows the user to define video objects on multiple frames and the system to interpolate the video object contours in every frame. Another video-browsing system is presented in Uchihashi, Foote, Girgensohn, and Boreczky (1999), where comic-book-style summaries are used to provide fast overviews of the video content. One other prototype retrieval system that supports 3D (three-dimensional) images, videos, and music retrieval is presented in Kosugi et al. (2001). In that system each type of query has its own processing module; for instance, image retrieval is processed using a component called ImageCompass.

Due to the importance of determining video similarity, a number of researchers have proposed various approaches to perform this task and a quick review follows.

In the context of image-retrieval systems, some researchers considered local geometric constraint into account and calculated the similarity between two images using the number of corresponding points (Lew, 2001). Others formulated the similarity between images as a graph-matching problem and used a graph-matching algorithm to calculate such similarity (Lew). In Oria, Ozsu, Lin, and Iglinski (2001) images are represented using a combination of color distribution (histogram) and salient objects (region of interest). Similarity between images is evaluated using a weighted Euclidean distance function, while complex query formulation was allowed using a modified version of SQL (structured query language) denoted as MOQL (multimedia object query language). Berretti, Bimbo, and Pala (2000) proposed a system that uses perceptual distance to measure the shape-feature similarity of images while providing efficient index structure.

One technique was proposed in Cheung and Zakhor (2000) that uses the metadata derived from clip links and the visual content of the clip to measure video similarity. At first, an abstract form of each video clip is calculated using a random set of images, then the closest frame in each video to a particular image in that set is found. The set of these closest frames is considered as a signature for that video clip. An extension to this work is introduced in Cheung and Zakhor (2001). In that article, the authors stated the need for a robust clustering algorithm to offset the errors produced by random sampling of the signature set. The clustering algorithm they proposed is based upon the graph theory. Another clustering algorithm was proposed in Liu, Zhuang, and Pan (1999) to dynamically distinguish whether two shots are similar or not based on the current situation of shot similarity.

A different retrieval approach uses time-alignment constraints to measure the similarity and dissimilarity of temporal documents. In Yamuna and Candan (2000), multimedia documents are viewed as a collection of objects linked to each other through various structures including temporal, spatial, and interaction structures. The similarity model in that work uses a highly structured class of linear constraints that is based on instant-based point formalism.

In Tan, Kulkarni, and Ramadge (1999), a framework is proposed to measure the video similarity. It employs different comparison resolutions for different phases of video search and uses color histograms to calculate frames similarity. Using this method, the evaluation of video similarity becomes equivalent to finding the path with the minimum cost in a lattice. In order to consider the temporal dimension of video streams without losing sight of the visual content, Adjeroh, Lee, and King (1999) considered the problem of video-stream matching as a pattern-matching problem. They devised the use of the vstring (video string) distance to measure video data similarity.

A powerful concept to improve searching multimedia databases is called relevance feedback (Wu, Zhuang, & Pan, 2000; Zhou & Huang, 2002). In this technique, the user associates a score to each of the returned hits, and these scores are used to direct the following search phase and improve its results. In Zhou and Huang, the authors defined relevance feedback as a biased classification problem in which there is an unknown number of classes but the user is only interested in one class. They used linear and nonlinear bias-discriminant analysis, which is a supervised learning scheme to solve the classification problem at hand. Brunelli and Mich (2000) introduced an approach that tunes search strategies and comparison metrics to user behavior in order to improve the effectiveness of relevance feedback.

EVALUATING VIDEO SIMILARITY USING A HUMAN-BASED MODEL

From the above survey of the current approaches, we can observe that an important issue has been overlooked by most of the above techniques. This was stated in Santini and Jain (1999, p. 882) by the following quote: "[I]f our systems have to respond in an intuitive and intelligent manner, they must use a similarity model resembling the humans." Our belief in the utmost importance of the above phrase motivates us to propose a novel technique to measure the similarity of video data. This approach attempts to introduce a model to emulate the way humans perceive video data similarity (Farag & Abdel-Wahab, 2003).

The retrieval system can accept queries in the form of an image, a single video shot, or a multishot video clip. The latter is the general case in video-retrieval systems. In order to lay the foundation of the proposed similarity-matching model, a number of assumptions are listed first.

- The similarity of video data (clip to clip) is based on the similarity of their constituent shots.
- Two shots are not relevant if the query signature (relative distance between selected key frames) is longer than the other signature.
- A database clip is relevant if one query shot is relevant to any of its shots.
- The query clip is usually much smaller than the average length of database clips.

The result of submitting a video clip as a search example is divided into two levels. The first one is the query overall similarity level, which lists similar database clips. In the second level, the system displays a list of similar shots to each shot of the input query, and this gives the user much more detailed results based on the similarity of individual shots to help fickle users in their decisions.

A shot is a sequence of frames, so we need first to formulate the first frames' similarity. In the proposed model, the similarity between two video frames is defined based on their visual content, where color and texture are used as visual content representative features. Color similarity is measured using the normalized histogram intersection, while texture similarity is calculated using a Gabor wavelet transform. Equation 1 is used to measure the overall similarity between two frames $f1$ and $f2$, where S_c (color similarity) is defined in Equation 2. A query frame histogram (H_{f1}) is scaled before applying Equation 2 to filter out variations in the video clips' dimensions. S_t (texture similarity) is calculated based on the mean and the standard deviation of each component of the Gabor filter (scale and orientation).

$$Sim(f1, f2) = 0.5 * S_c + 0.5 * S_t \quad (1)$$

$$S_c = \left[\sum_{i=1}^{64} Min(H_{f1}(i), H_{f2}(i)) \right] / \sum_{i=1}^{64} H_{f1}(i) \quad (2)$$

Suppose we have two shots $S1$ and $S2$, and each has $n1$ and $n2$ frames respectively. We measure the similarity between these shots by measuring the similarity between every frame in $S1$ with every frame in $S2$, and form what we call the similarity matrix that has a dimension of $n1$x$n2$. For the ith row of the similarity matrix, the largest element value represents the closest frame in shot $S2$ that is most similar to the ith frame in shot $S1$ and vice versa. After forming that matrix, Equation 3 is used to measure shot similarity. Equation 3 is applied upon the selected key frames to improve efficiency and avoid redundant operations.

$$Sim(S1, S2) = \left[\sum_{i=1}^{n1} MR_{(i)}(S_{i,j}) + \sum_{j=1}^{n2} MC_{(j)}(S_{i,j}) \right] / (n1 + n2) \quad (3)$$

where $MR_{(i)}(S_{i,j})/MC_{(j)}(S_{i,j})$ is the element with the maximum value in the i/j row and column respectively, and $n1/n2$ is the number of rows and columns in the similarity matrix.

The proposed similarity model attempts to emulate the way humans perceive the similarity of video material. This was achieved by integrating into the similarity-measuring Equation 4 a number of factors that humans most probably use to perceive video similarity. These factors are the following.

- **The visual similarity:** Usually humans determine the similarity of video data based on their visual characteristics such as color, texture, shape, and so forth. For instance, two images

with the same colors are usually judged as being similar.

- **The rate of playing the video:** Humans tend also to be affected by the rate at which frames are displayed, and they use this factor in determining video similarity.
- **The time period of the shot:** The more the periods of video shots coincide, the more they are similar to human perception.
- **The order of the shots in a video clip:** Humans often give higher similarity scores to video clips that have the same ordering of corresponding shots.

$$Sim(S1,S2) \quad W_1 * S_V \quad W_2 * D_R \quad W_3 * F_R \tag{4}$$

$$D_R \quad 1 \quad |S1(d) \quad S2(d)| / Max(S1(d), S2(d)) \tag{5}$$

$$F_R \quad 1 \quad |S1(r) \quad S2(r)| / Max(S1(r), S2(r)) \tag{6}$$

where S_V is the visual similarity, D_R is the shot-duration ratio, F_R is the video frame-rate ratio, $Si(d)$ is the time duration of the ith shot, $Si(r)$ is the frame rate of the ith shot, and W_1, W_2, and W_3 are relative weights.

There are three parameter weights in Equation 4, namely, W_1, W_2, and W_3, that give indication on how important a factor is over the others. For example, stressing the importance of the visual similarity factor is achieved by increasing the value of its associated weight (W_1). It was decided to give the user the ability to express his or her real need by allowing these parameters to be adjusted by the user. To reflect the effect of the order factor, the overall similarity level checks if the shots in the database clip have the same temporal order as those shots in the query clip. Although this may restrict the candidates to the overall similarity set to clips that have the same temporal order of shots as the query clip, the user still has a finer level of similarity that is based on individual query shots, which captures other aspects of similarity as discussed before.

To evaluate the proposed similarity model, it was implemented in the retrieval stage of the VCR system (a video content-based retrieval system). The model performance was quantified through measuring recall and precision defined in Equations 7 and 8. To measure the recall and precision of the system, five shots were submitted as queries while the returned-shots number was changed from five to 20. Both recall and precision depend on the number of returned shots. To increase recall, more shots have to be retrieved, which will in general result in a decreased precision. The average recall and precision is calculated for the above experiments and plotted in Figure 1, which indicates a very good performance achieved by the system. At a small number of returned shots, the recall value was small while the precision value was very good. Increasing the number of returned clips increases the recall until it reaches one; at the same time the value of the precision was not degraded very much, but the curve almost dwells at a precision value of 0.92. This way, the system provides a very good trade-off between recall and precision. Similar results were obtained using the same procedure for unseen queries. For more discussion on the obtained results, the reader is referred to Farag and Abdel-Wahab (2003).

$$R \quad A/(A \quad C) \tag{7}$$

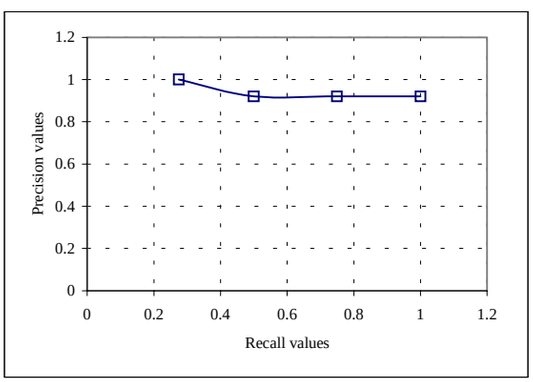
$$P \quad A/(A \quad B) \tag{8}$$

A: correctly retrieved, B: incorrectly retrieved, C: missed

FUTURE TRENDS

The proposed model is one step to solve the problem of modeling human perception in measuring video data similarity. Many open research topics and outstanding problems still exit, and a brief review follows. Since Euclidean measure may not effectively emulate human perception, the potential of improving it can be explored via clustering and

Figure 1. Recall vs. precision for five seen shots



neural-network techniques. Also, there is a need to propose techniques that measure the attentive similarity, which is what humans actually use while judging multimedia data similarity. Moreover, nonlinear methods for combining more than one similarity measure require more exploration. The investigation of methodologies for performance evaluation of multimedia retrieval systems and the introduction of benchmarks are other areas that need more research. In addition, semantic-based retrieval and how to correlate semantic objects with low-level features is another open topic. Finally, the introduction of new psychological similarity models that better capture the human notion of multimedia similarity is an issue that needs further investigation.

CONCLUSION

In this article, a brief introduction to the issue of measuring digital video data similarity is introduced in the context of designing effective content-based video-retrieval systems. The utmost significance of the similarity-matching model in determining the applicability and effectiveness of the retrieval system was emphasized. Afterward, the article reviewed some of the techniques proposed by the research community to implement the retrieval stage in general and to tackle the problem of assessing the similarity of multimedia data in particular. The proposed similarity-matching model is then introduced. This novel model attempts to measure the similarity of video data based on a number of factors that most probably reflect the way humans judge video similarity. The proposed model is considered a step on the road toward appropriately modeling the human's notion of multimedia data similarity. There are still many research topics and open areas that need further investigation in order to come up with better and more effective similarity-matching techniques.

REFERENCES

Adjeroh, D., Lee, M., & King, I. (1999). A distance measure for video sequences. *Journal of Computer Vision and Image Understanding, 75*(1/2), 25-45.

Berretti, S., Bimbo, A., & Pala, P. (2000). Retrieval by shape similarity with perceptual distance and effective indexing. *IEEE Transactions on Multimedia*, *2*(4), 225-239.

Brunelli, R., & Mich, O. (2000). Image retrieval by examples. *IEEE Transactions on Multimedia*, *2*(3), 164-171.

Brunelli, R., Mich, O., & Modena, C. (1999). A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, *10*(2), 78-112.

Cheung, S., & Zakhor, A. (2000). Efficient video similarity measurement and search. *Proceedings of IEEE International Conference on Image Processing*, (pp. 85-89).

Cheung, S., & Zakhor, A. (2001). Video similarity detection with video signature clustering. *Proceedings of IEEE International Conference on Image Processing*, (pp. 649-652).

Deb, S. (2004). *Multimedia systems and content-based retrieval*. Hershey, PA: Idea Group Publishing.

Farag, W., & Abdel-Wahab, H. (2001). A new paradigm for detecting scene changes on MPEG compressed videos. *Proceedings of IEEE International Symposium on Signal Processing and Information Technology*, (pp. 153-158).

Farag, W., & Abdel-Wahab, H. (2002a). Adaptive key frames selection algorithms for summarizing video data. *Proceedings of the Sixth Joint Conference on Information Sciences*, (pp. 1017-1020).

Farag, W., & Abdel-Wahab, H. (2002b). A new paradigm for analysis of MPEG compressed videos. *Journal of Network and Computer Applications*, *25*(2), 109-127.

Farag, W., & Abdel-Wahab, H. (2003). A human-based technique for measuring video data similarity. *Proceedings of the Eighth IEEE International Symposium on Computers and Communications (ISCC2003)*, (pp. 769-774).

Kosugi, N., Nishimura, G., Teramoto, J., Mii, K., Onizuka, M., Kon'ya, S., et al. (2001). Content-based retrieval applications on a common database

management system. *Proceedings of ACM International Conference on Multimedia*, (pp. 599-600).

Lew, M. (Ed.). (2001). *Principles of visual information retrieval*. London: Springer-Verlag.

Liu, X., Zhuang, Y., & Pan, Y. (1999). A new approach to retrieve video by example video clip. *Proceedings of ACM International Conference on Multimedia*, (pp. 41-44).

Luo, H., & Eleftheriadis, A. (1999). Designing an interactive tool for video object segmentation and annotation: Demo abstract. *Proceedings of ACM International Conference on Multimedia*, (p. 196).

Oria, V., Ozsu, M., Lin, S., & Iglinski, P. (2001). Similarity queries in DISIMA DBMS. *Proceedings of ACM International Conference on Multimedia*, (pp. 475-478).

Santini, S., & Jain, R. (1999). Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(9), 871-883.

Tan, Y., Kulkarni, S., & Ramadge, P. (1999). A framework for measuring video similarity and its application to video query by example. *Proceedings of IEEE International Conference on Image Processing*, (pp. 106-110).

Uchihashi, S., Foote, J., Girgensohn, A., & Boreczky, J. (1999). Video manga: Generating semantically meaningful video summaries. *Proceedings of ACM International Conference on Multimedia*, (pp. 383-392).

Wu, Y., Zhuang, Y., & Pan, Y. (2000). Content-based video similarity model. *Proceedings of ACM International Conference on Multimedia*, (pp. 465-467).

Yamuna, P., & Candan, K. (2000). Similarity-based retrieval of temporal documents. *Proceedings of ACM International Conference on Multimedia*, (pp. 243-246).

Yoshitaka, A., & Ichikawa, T. (1999). A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, *11*(1), 81-93.

Zhou, X., & Huang, T. (2002). Relevance feedback in content-based image retrieval: Some recent advances. *Proceedings of the Sixth Joint Conference on Information Sciences*, (pp. 15-18).

KEY TERMS

Color Histogram: A method to represent the color feature of an image by counting how many values of each color occur in the image, and then form a representing histogram.

Content-Based Access: A technique that enables searching multimedia databases based on the content of the medium itself and not based on a keywords description.

Multimedia Databases: Nonconventional databases that store various media such as images and audio and video streams.

Query by Example: A technique to query multimedia databases in which the user submits a sample query and asks the system to retrieve similar items.

Relevance Feedback: A technique in which the user associates a score to each of the returned hits, then these scores are used to direct the following search phase and improve its results.

Retrieval Stage: The last stage in a content-based retrieval system that accepts and processes a user query, then returns the results ranked according to their similarities with the query.

Similarity Matching: The process of comparing extracted features from the query with those stored in the metadata.

Asymmetric Digital Subscriber Line

Leo Tan Wee Hin
Singapore National Academy of Science and Nanyang Technological University, Singapore

R. Subramaniam
Singapore National Academy of Science and Nanyang Technological University, Singapore

INTRODUCTION

The plain, old telephone system (POTS) has formed the backbone of the communications world since its inception in the 1880s. Running on twisted pairs of copper wires bundled together, there has not really been any seminal developments in its mode of transmission, save for its transition from analogue to digital toward the end of the 1970s.

The voice portion of the line, including the dial tone and ringing sound, occupies a bandwidth that represents about 0.3% of the total bandwidth of the copper wires. This seems to be such a waste of resources, as prior to the advent of the Internet, telecommunication companies (telcos) have not really sought to explore better utilization of the bandwidth through technological improvements, for example, to promote better voice quality, to reduce wiring by routing two neighboring houses on the same line before splitting the last few meters, and so on. There could be two possible reasons for this state of affairs. One reason is that the advances in microelectronics and signal processing necessary for the efficient and cost-effective interlinking of computers to the telecommunications network have been rather slow (Reusens, van Bruyssel, Sevenhans, van Den Bergh, van Nimmen, & Spruyt, 2001). Another reason is that up to about the 1990s, telcos were basically state-run enterprises that had little incentive to roll out innovative services and applications. When deregulation and liberalization of the telecommunication sector was introduced around the 1990s, the entire landscape underwent a drastic transformation and saw telcos introducing a plethora of service enhancements, innovations, and other applications; there was also a parallel surge in technological developments aiding these.

As POTS is conspicuous by its ubiquity, it makes sense to leverage on it for upgrading purposes rather than deploy totally new networks that need considerable investment. In recent times, asymmetric digital subscriber line (ADSL) has emerged as a technology that is revolutionizing telecommunications and is a prime candidate for broadband access to the Internet. It allows for the transmission of enormous amounts of digital information at rapid rates on the POTS.

BACKGROUND

The genesis of ADSL can be traced to the efforts by telecommunication companies to enter the cable-television market (Reusens et al., 2001). They were looking for a way to send television signals over the phone line so that subscribers can also use this line for receiving video.

The foundations of ADSL were laid in 1989 by Joseph Leichleder, a scientist at Bellcore, who observed that there are more applications and services for which speedier transmission rates are needed from the telephone exchange to the subscriber's location than for the other way around (Leichleider, 1989). Telcos working on the video-on-demand market were quick to recognize the potential of ADSL for streaming video signals. However, the video-on-demand market did not take off for various reasons: Telcos were reluctant to invest in the necessary video architecture as well as to upgrade their networks, the quality of the MPEG (Moving Picture Experts Group) video stream was rather poor, and there was competition from video rental stores (Reusens et al., 2001). Also, the hybrid fiber coaxial (HFC) architecture for cable television, which was introduced around 1993, posed a serious challenge. At about this time, the Internet was becoming a phenomenon, and telcos were quick to realize the potential of ADSL for fast Internet access. Field trials began in 1996, and in 1998, ADSL started to be deployed in many countries.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

The current motivation of telcos in warming toward ADSL has more to do with the fact that it offers rapid access to the Internet, as well as the scope to deliver other applications and services whilst offering competition to cable-television companies entering the Internet-access market. All this means multiple revenue streams for telcos.

Over the years, technological advancements relating to ADSL as well as the evolution of standards for its use have begun to fuel its widespread deployment for Internet access (Chen, 1999). Indeed, it is one of those few technologies that went from the conceptual stage to the deployment stage within a decade (Starr, Cioffi, & Silverman, 1999).

This article provides an overview of ADSL.

ADSL TECHNOLOGY

ADSL is based on the observation that while the frequency band for voice transmission over the phone line occupies about 3 KHz (200 Hz to 3,300 Hz), the actual bandwidth of the twisted pairs of copper wires constituting the phone line is more than 1 MHz (Hamill, Delaney, Furlong, Gantley, & Gardiner, 1999; Hawley, 1999). It is the unused bandwidth beyond the voice portion of the phone line that ADSL uses for transmitting information at high rates. A high frequency (above 4,000 KHz) is used because more information can then be transmitted at faster rates; a disadvantage is that the signals undergo attenuation with distance, which restricts the reach of ADSL.

There are three key technologies involved in ADSL.

Signal Modulation

Modulation is the process of transmitting information on a wire after encoding it electrically. When ADSL was first deployed on a commercial basis, carrierless amplitude-phase (CAP) modulation was used to modulate signals over the line. CAP works by dividing the line into three subchannels: one for voice, one for upstream access, and another for downstream access. It has since been largely superseded by another technique called discrete multitone (DMT), which is a signal-coding technique invented by John Cioffi of Stanford University (Cioffi, Silverman, & Starr, 1999; Ruiz, Cioffi, & Kasturia, 1992). He demonstrated its use by transmitting 8 Mb of information in one second across a phone line 1.6 km long. DMT scores over CAP in terms of the speed of data transfer, efficiency of bandwidth allocation, and power consumption, and these have been key considerations in its widespread adoption.

DMT divides the bandwidth of the phone line into 256 subchannels through a process called frequency-division multiplexing (FDM; Figure 1; Kwok, 1999). Each subchannel occupies a bandwidth of 4.3125 KHz. For transmitting data across each subchannel, the technique of quadrature amplitude modulation (QAM) is used. Two sinusoidal carriers of the same frequency that differ in phase by 90 degrees constitute the QAM signal. The number of bits allocated for each subchannel varies from 2 to 16: Higher bits are carried on subchannels in the lower frequencies, while lower bits are carried on channels in the higher frequencies.

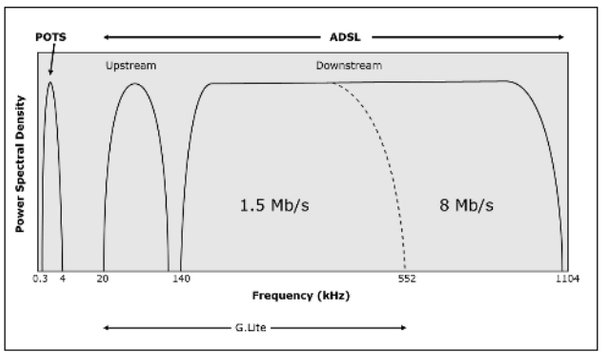
The following theoretical rates apply.

- **Upstream access:**
  20 carriers x 8 bits x 4 KHz = 640 Kbps
- **Downstream access:**
  256 carriers x 8 bits x 4 KHz = 8.1 Mbps

In practice, the data rates achieved are much less owing to inadequate line quality, extended length of line, cross talk, and noise (Cook, Kirkby, Booth, Foster, Clarke, & Young, 1999). The speed for downstream access is generally about 10 times that for upstream access.

Two of the channels (16 and 64) can be used for transmitting pilot signals for specific applications or tests. It is the subdivision into 256 channels that allows one group to be used for downstream access and another for upstream access on an optimal basis. When the modem is activated during network access, the signal-to-noise ratio in the channel is automatically measured. Subchannels that experience unacceptable throughput of the signal owing to interference are turned off, and their traffic is redirected to other suitable subchannels, thus optimizing the overall transmission throughput. The total transmittance is thus maintained by QAM. This is a particular advantage when using POTS for ADSL delivery since a good portion of the network was laid several decades ago and is susceptible to interference owing to corrosion and other problems.

Figure 1. Frequency-division multiplexing of ADSL



The upstream channel is used for data transmission from the subscriber to the telephone exchange, while the downstream channel is used for the converse link. It is this asymmetry in transmission rates that accounts for the asymmetry in ADSL. As can be seen from Figure 1, the voice portion of the line is separated from the data-transmission portion; this is accomplished through the use of a splitter. It is thus clear why phone calls can be made over the ADSL link even during Internet access. At frequencies where the upstream and downstream channels need to overlap for part of the downstream transmission, so as to make better use of the lower frequency region where signal loss is less, the use of echo-cancellation techniques is necessary to ensure the differentiation of the mode of signal transmission (Winch, 1998).

Code and Error Correction

The fidelity of information transmitted on the phone line is contingent on it being coded suitably and decoded correctly at the destination even if some bits of information are lost during transmission. This is commonly accomplished by the use of constellation encoding and decoding. Further enhancements in reliability is afforded by a technique called forward error correction, which minimizes the necessity for retransmission (Gillespie, 2001).

Framing and Scrambling

The effectiveness of coding and error correction is greatly enhanced by sequentially scrambling the data. To accomplish this, the ADSL terminal unit at the control office (ATU-C) transmits 68 data frames every 17 ms, with each of these data frames obtaining its information from two data buffers (Gillespie, 2001).

STANDARDS FOR ADSL

The deployment of ADSL has been greatly facilitated by the evolution of standards laid by various international agencies. These standards are set after getting input from carriers, subscribers, and service providers. The standards dictate the operation of ADSL under a variety of conditions and cover aspects such as equipment specifications, connection protocols, and transmission metrics (Chen, 1999; Summers, 1999). The more important of these standards are indicated below.

- **G.dmt:** Also known as full-rate ADSL or G992.1, it is the first version of ADSL.
- **G.Lite:** Also known as universal ADSL or G992.2, it is the standard method for installing ADSL without the use of splitters. It permits downstream access at up to 1.5 Mbps and upstream access at up to 512 Kbps over a distance of up to 18,000 ft.
- **ADSL2:** Also known as G 992.3 and G 992.4, it is a next-generation version that allows for even higher rates of data transmission and extension of reach by 180 m.
- **ADSL2+:** Also known as G 992.5, this variant of ADSL2 doubles the speed of transmission of signals from 1.1 MHz to 2.2 MHz, as well as extends the reach even further.
- **T1.413:** This is the standard for ADSL used by the American National Standards Institute (ANSI), and it depends on DMT for signal modulation. It can achieve speeds of up to 8 Mbps for downstream access and up to 1.5 Mbps for upstream access over a distance of 9,000 to 12,000 ft.
- **DTR/TM-06001:** This is an ADSL standard used by the European Technical Standards Institute (ETSI) and is based on T1.413, but modified to suit European conditions

The evolution of the various ADSL variants is a reflection of the technological improvements that have occurred in tandem with the increase in subscriber numbers.

OPERATIONAL ASPECTS

Where the telephone exchange has been ADSL enabled, setting up the ADSL connection for a subscriber is a straightforward task. The local loop from the subscriber's location is first linked via a splitter to the equipment at the telephone exchange, and an ADSL modem is then interfaced to the loop at this exchange. Next, a splitter is affixed to the telephone socket at the subscriber's location, and the lead wire from the phone is linked to the rear of the splitter and an ADSL modem. The splitters separate the telephony signal from the data streams, while the modems at the telephone exchange and subscriber location cater for the downstream and upstream data flow, respectively. A network device called digital subscriber line access multiplexer (DSLAM) at the exchange splits signals from subscriber lines into two streams: The voice portion is carried on POTS while the data portion is fed to a high-speed backbone using multiplexing techniques and then to the Internet (Green, 2001). A schematic of the ADSL setup is illustrated in Figure 2.

The installation of the splitter at the subscriber's premises is a labor-intensive task as it requires a technician to come and do the necessary work. This comes in the way of widespread deployment of ADSL by telcos. A variant of ADSL known as splitterless ADSL (G992.2) or G.Lite was thus introduced to address this (Kwok, 1999).

Speeds attainable on an ADSL link are variable and are higher than that obtained using a 56-K modem. The speed is also distance dependent (Table 1; Azzam & Ransom, 1999). This is because the high frequency signals undergo attenuation with distance and, as a result, the bit rates transmitted via the modem decrease accordingly. Other factors that can affect the speed include the quality of the copper cables, the extent of network congestion, and, for overseas access, the amount of international bandwidth leased by the Internet service providers (ISPs). The latter factor is not commonly recognized.

ADVANTAGES AND DISADVANTAGES OF ADSL

Any new technology is not perfect, and there are constraints that preclude its optimal use; this has to be addressed by ongoing research and development. The following are some of the advantages of ADSL.

- It does not require the use of a second phone line.
- It can be installed on demand, unlike fiber cabling, which requires substantial underground work as well as significant installation work at the subscriber's location.
- It provides affordable broadband access at speeds significantly greater than that obtainable using a dial-up modem.

Figure 2. Architecture of ADSL (G992.1) setup

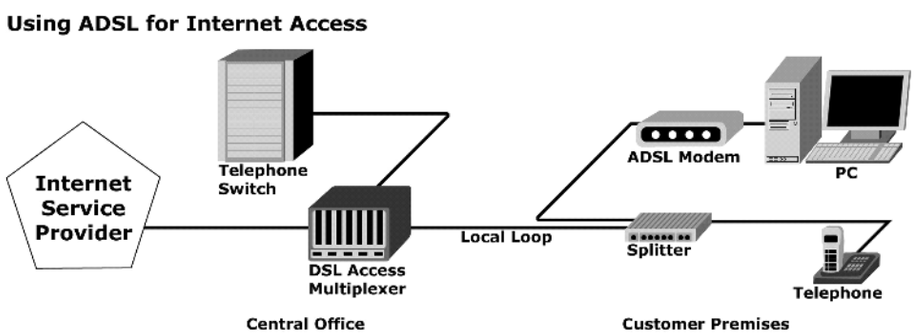


Table 1. Performance of ADSL

ADSL Wire Gauge (AWG)	Distance (ft)	Upstream Rate (Kbps)	Downstream Rate (Mbps)
24	18,000	176	1.7
26	13,500	176	1.7
24	12,000	640	6.8
26	9,000	640	6.8

- Since there is a dedicated link between the subscriber's location and the telephone exchange, there is greater security of the data as compared to other alternatives such as cable modem.
- No dial up is needed as the connection is always on.

Some of the disadvantages of ADSL are as follows.

- The subscriber's location needs to be within about 5 km from the telephone exchange; the greater the distance away from the exchange, the less is the speed of data transfer.
- As ADSL relies on copper wires, a good proportion of which was laid underground and overland many years ago, the line is susceptible to noise due to, for example, moisture, corrosion, and cross talk, all of which can affect its performance (Cook, Kirkby, Booth, Foster, Clarke & Young, 1999).

On the balance, the advantages of ADSL far outweigh its disadvantages, and this has led to its deployment in many countries for broadband access, for example, in Singapore (Tan & Subramaniam, 2000, 2001).

APPLICATIONS

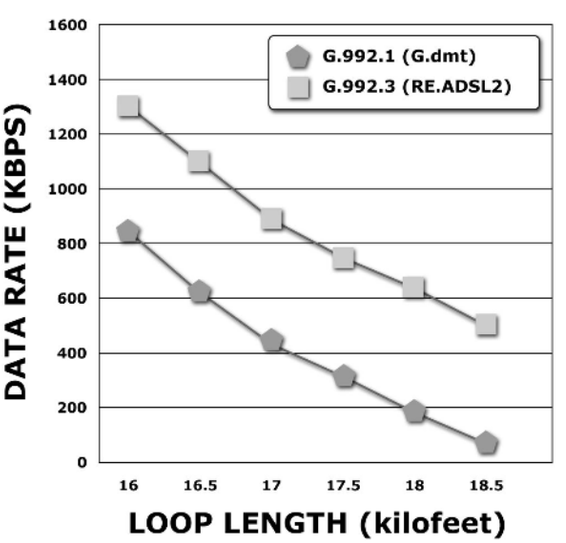
Currently, ADSL is used mainly for broadband access, that is, for high-speed Internet access as well as for rapid downloading of large files. Other applications include accessing video catalogues, image libraries (Stone, 1999), and digital video libraries (Smith, 1999); playing interactive games that guzzle bandwidth; accessing remote CD-ROMs; videoconferencing; distance learning; network computing whereby software and files can be stored in a central server and then retrieved at fast speeds (Chen, 1999); and telemedicine, in which patients can access specialist expertise in remote locations for real-time diagnostic advice, which may include the examination of high-quality X-ray films and other biomedical images.

Future applications could include television, Internet telephony, and other interactive applications, all of which can lead to increase in revenue for telcos. There is a possibility that video-on-demand can take off.

FUTURE TRENDS

The maturation of ADSL is being fueled by technological advances. The number of subscribers for ADSL has seen an upward trend in many countries (Kalakota, Gundepudi, Wareham, Rai, & Weike, 2002). New developments in DMT are likely to lead to more efficient transmission of data streams. The distance-dependent nature of its transmission is likely to be overcome either by the building of more telephone exchanges so that more subscriber locations can be within an effective radius for the deployment of ADSL, or by advances in the enabling technologies. The technology is likely to become more pervasive than its competitor, cable modem, in the years to come since the installation of new cabling will take

Figure 3. Comparison of two ADSL variants



years to reach more households and also entails further investments.

The higher variants of ADSL such as ADSL2 and ADSL2+ are likely to fuel penetration rates further (Tzannes, 2003). For example, compared to first-generation ADSL, ADSL2 can enhance data rates by 50 Kbps and reach by 600 ft (Figure 3), the latter translating to an increase in area coverage by about 5%, thus raising the prospects of bringing more subscribers on board.

Some of the features available with the new variants of ADSL, such as automatic monitoring of line quality and signal-to-noise ratio, offers the potential to customize enhanced service-delivery packages at higher tariffs for customers who want a higher quality of service.

CONCLUSION

Twisted pairs of copper wires forming the POTS constitute the most widely deployed access network for telecommunications. Since ADSL leverages on the ubiquity of this network, it allows telcos to extract further mileage without much additional investments whilst competing with providers of alternative platforms. It is thus likely to be a key broadband technology for Internet access in many countries in the years to come. A slew of applications that leverage on ADSL are also likely to act as drivers for its widespread deployment.

ACKNOWLEDGEMENTS

We thank Dr. Tan Seng Chee and Mr. Derrick Yuen for their assistance with the figures.

REFERENCES

Azzam, A., & Ransom, N. (1999). *Broadband access technologies: ADSL/VDSL, cable modems, fiber, LMDS.* New York: McGraw-Hill.

Chen, W. (1999, May). The development and standardization of asymmetric digital subscriber lines. *IEEE Communications Magazine, 37*(5), 68-70.

Cioffi, J., Silverman, P., & Starr, T. (1998, December). Digital subscriber lines. *Computer Networks, 31*(4), 283-311.

Cook, J., Kirkby, R., Booth, M., Foster, K., Clarke, D., & Young, G. (1999). The noise and crosstalk environments for ADSL and VDSL systems. *IEEE Communications Magazine*, *37*(5), 73-78.

Gillespie, A. (2001). *Broadband access technologies, interfaces and management.* Boston: Artech House.

Green, J. H. (2001). *The Irwin handbook of telecommunications.* New York: McGraw-Hill.

Hamill, H., Delaney, C., Furlong, E., Gantley, K., & Gardiner, K. (1999). *Asymmetric digital subscriber line.* Retrieved August 15, 2004, from *http://www.esatclear.ie/~aodhoh/adsl/report.html*

Hawley, G. T. (1999). DSL: Broadband by phone. *Scientific American, 281,* 82-83.

Kalakota, R., Gundepudi, P., Wareham, J., Rai, A., & Weike, R. (2002). The economics of DSL regulation. *IEEE Computer Magazine, 35*(10), 29-36.

Kwok, T. C. (1999, May). Residential broadband architecture over ADSL and G.lite (G992.4): PPP over ATM. *IEEE Communications Magazine,* 37(5), 84-89.

Lechleider, J. L. (1989, September 4). *Asymmetric digital subscriber lines* [Memo]. NJ: Bell Communication Research.

Reusens, P., van Bruyssel, D., Sevenhans, J., van Den Bergh, S., van Nimmen, B., & Spruyt, P. (2001). A practical ADSL technology following a decade of effort. *IEEE Communications Magazine, 39*(1), 145-151.

Ruiz, A., Cioffi, I. M., & Kasturia, S. (1992). Discrete multi tone modulation with coset coding for the spectrally shaped channel. *IEEE Transactions, 40*(6), 1012-1027.

Smith, J. R. (1999). Digital video libraries and the Internet. *IEEE Communications Magazine, 37*(1), 92-97.

Starr, T., Cioffi, J. M., & Silverman, P. J. (1999). *Understanding digital subscriber line technology.* New York: Prentice Hall.

Stone, H. S. (1999). Image libraries and the Internet. *IEEE Communications Magazine, 37*(1), 99-106.

Summers, C. (1999). *ADSL standards implementation and architecture*. Boca Raton, FL: CRC Press.

Tan, W. H. L., & Subramaniam, R. (2000). Wiring up the island state. *Science, 288*, 621-623.

Tan, W. H. L., & Subramaniam, R. (2001). ADSL, HFC and ATM technologies for a nationwide broadband network. In N. Barr (Ed.), *Global communications 2001* (pp. 97-102). London: Hanson Cooke Publishers.

Tzannes, M. (2003). *RE-ADSL2: Helping extend ADSL's reach.* Retrieved September 15, 2004, from *http://www.commsdesign.com/design_library/cd/hn/OEG20030513S0014*

Winch, R. G. (1998). *Telecommunication transmission systems*. New York: McGraw-Hill.

KEY TERMS

ADSL: Standing for asymmetric digital subscriber line, it is a technique for transmitting large amounts of data rapidly on twisted pairs of copper wires, with the transmission rates for downstream access being much greater than for the upstream access.

Bandwidth: Defining the capacity of a communication channel, it refers to the amount of data that can be transmitted in a fixed time over the channel; it is commonly expressed in bits per second.

Broadband Access: This is the process of using ADSL, fiber cable, or other technologies to transmit large amounts of data at rapid rates.

CAP: Standing for carrierless amplitude-phase modulation, it is a modulation technique in which the entire frequency range of a communications line is treated as a single channel and data is transmitted optimally.

DMT: Standing for discrete multitone technology, it is a technique for subdividing a transmission channel into 256 subchannels of different frequencies through which traffic is overlaid.

Forward Error Correction: It is a technique used in the receiving system for correcting errors in data transmission.

Frequency-Division Multiplexing: This is the process of subdividing a telecommunications line into multiple channels, with each channel allocated a portion of the frequency of the line.

Modem: This is a device that is used to transmit and receive digital data over a telecommunications line.

MPEG: This is an acronym for Moving Picture Experts Group and refers to the standards developed for the coded representation of digital audio and video.

QAM: Standing for quadrature amplitude modulation, it is a modulation technique in which two sinusoidal carriers that have a phase difference of 90 degrees are used to transmit data over a channel, thus doubling its bandwidth.

SNR: Standing for signal-to-noise ratio, it is a measure of signal integrity with respect to the background noise in a communication channel.

Splitter: This is a device used to separate the telephony signals from the data stream in a communications link.

Twisted Pairs: This refers to two pairs of insulated copper wires intertwined together to form a communication medium.

ATM Technology and E-Learning Initiatives

Marlyn Kemper Littman
Nova Southeastern University, USA

INTRODUCTION

The remarkable popularity of Web-based applications featuring text, voice, still images, animations, full-motion video and/or graphics and spiraling demand for broadband technologies that provision seamless multimedia delivery motivate implementation of asynchronous transfer mode (ATM) in an array of electronic learning (e-learning) environments (Parr & Curran, 2000). Asynchronous refers to ATM capabilities in supporting intermittent bit rates and traffic patterns in response to actual demand, and transfer mode indicates ATM capabilities in transporting multiple types of network traffic.

E-learning describes instructional situations in which teachers and students are physically separated (Lee, Hou & Lee, 2003; Hunter & Carr, 2002). ATM is a high-speed, high-performance multiplexing and switching communications technology that bridges the space between instructors and learners by providing bandwidth on demand for enabling interactive real-time communications services and delivery of multimedia instructional materials with quality-of-service (QoS) guarantees.

Research trials and full-scale ATM implementations in K-12 schools and post-secondary institutions conducted since the 1990s demonstrate this technology's versatility in enabling telementoring, telecollaborative research and access to e-learning enrichment courses. However, with enormous bandwidth provided via high-capacity 10 Gigabit Ethernet, wavelength division multiplexing (WDM) and dense WDM (DWDM) backbone networks; high costs of ATM equipment and service contracts; and interoperability problems between different generations of ATM core components such as switches, ATM is no longer regarded as a universal broadband solution.

Despite technical and financial issues, ATM networks continue to support on-demand access to Web-based course content and multimedia applications. ATM implementations facilitate the seamless integration of diverse network components that include computer systems, servers, middleware, Web caches, courseware tools, digital library materials and instructional resources such as streaming video clips in dynamic e-learning system environments. National research and education networks (NRENs) in countries that include Belgium, Croatia, Estonia, Greece, Israel, Latvia, Moldavia, Portugal, Spain, Switzerland and Turkey use ATM in conjunction with technologies such as Internet protocol (IP), synchronous digital hierarchy (SDH), WDM and DWDM in supporting synchronous and asynchronous collaboration, scientific investigations and e-learning initiatives (TERENA, 2003).

This article reviews major research initiatives contributing to ATM development. ATM technical fundamentals and representative ATM specifications are described. Capabilities of ATM technology in supporting e-learning applications and solutions are examined. Finally, trends in ATM implementation are explored.

BACKGROUND

Bell Labs initiated work on ATM research projects during the 1960s and subsequently developed cell switching architecture for transporting bursty network traffic. Initially known as asynchronous time-division multiplexing (ATDM), ATM was originally viewed as a replacement for the time-division multiplexing (TDM) protocol that supported transmission of time-dependent and time-independent traffic and assigned each fixed-sized packet or cell to a fixed timeslot for transmission. In contrast to TDM, the ATM protocol dynamically allocated timeslots to cells on demand to accommodate application requirements.

In the 1990s, the foundation for practical ATM e-learning implementations was established in the European Union (EU) with the Joint ATM Experiment on European Services (JAMES); Trans-European

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Network-34.368 Mbps or megabits per second (TEN-34); and TEN-155.52 Mbps (TEN-155) projects. EU NRENs such as Super Joint Academic Network (SuperJANET) in the United Kingdom and SURFnet in The Netherlands demonstrated ATMs' dependable support of multimedia applications with QoS guarantees, interactive videoconferences and IP multicasts via optical connections at rates reaching 2.488 gigabits per second (Gbps, or in OC-48 in terms of optical carrier levels).

Implemented between 1994 and 1999, the European Commission (EC) Advanced Communications Technology and Services (ACTS) Program demonstrated ATM technical capabilities in interworking with wireline and wireless implementations. For instance, the EC ACTS COIAS (convergence of Internet ATM satellite) project confirmed the use of IP version 6 (IPv6) in enhancing network functions in hybrid satellite and ATM networks. The EC ACTS AMUSE initiative validated ATM-over-asynchronous digital subscriber line (ADSL) capabilities in delivering time-critical interactive broadband services to residential users (Di Concetto, Pavarani, Rosa, Rossi, Paul & Di Martino, 1999).

A successor to the EC ACTS Program, the EC Community Research and Development Information Service (CORDIS) Fifth Framework Information Society Technologies (IST) Program sponsored technical initiatives in the ATM arena between 1998 and 2002. For example, the open platform for enhanced interactive services (OPENISE) project verified capabilities of the ATM platform in interworking with ADSL and ADSL.Lite in supporting multimedia services and voice-over-ATM implementations. The creation and deployment of end user services in premium IP networks (CADENUS) initiative confirmed the effectiveness of ATM, IP and multiprotocol label switching (MPLS) operations in facilitating delivery of multimedia applications with QoS guarantees via mixed-mode wireline and wireless platform. The IASON (generic evaluation platform for services interoperability and networks) project validated the use of ATM in conjunction with an array of wireline and wireless technologies including universal mobile telecommunications systems (UMTS), IP, integrated services digital network (ISDN) and general packet radio service (GPRS) technologies. The WINMAN (WDM and IP network management) initiative demonstrated ATM, SDH and DWDM support of reliable IP transport and IP operations in conjunction with flexible and extendible network architectures. The NETAGE (advanced network adapter for the new generation of mobile and IP-based networks) initiative verified ATM, ISDN and IP functions in interworking with global systems for mobile communications (GSM), a 2G (second generation) cellular solution, and GPRS implementations.

Research findings from the Fifth Framework Program also contributed to the design of the transborder e-learning initiative sponsored by the EC. Based on integrated information and communications technology (ICT), this initiative supports advanced e-learning applications that respect language and cultural diversity and promotes digital literacy, telecollaborative research, professional development and lifelong education.

In the United States (U.S.), an IP-over-ATM-over-synchronous optical network (SONET) infrastructure served as the platform for the very high-speed broadband network service (vBNS) and its successor vBNS+, one of the two backbone networks that originally provided connections to Internet2 (I2). A next-generation research and education network, sponsored by the University Consortium for Advanced Internet Development (UCAID), I2 supports advanced research and distance education applications with QoS guarantees. Although replacement of ATM with ultra-fast DWDM technology as the I2 network core is under way, ATM technology continues to provision multimedia services at I2 member institutions that include the Universities of Michigan, Mississippi and Southern Mississippi, and Northeastern and Mississippi State Universities.

ATM TECHNICAL FUNDAMENTALS

To achieve fast transmission rates, ATM uses a standard fixed-sized 53-byte cell featuring a 5-byte header or addressing and routing mechanism that contains a virtual channel identifier (VCI), a virtual path indicator (VPI) and an error-detection field and a 48-byte payload or information field for transmission. ATM supports operations over physical media that include twisted copper wire pair and optical fiber with optical rates at 13.27 Gbps (OC-192). Since ATM enables connection-oriented services, information is transported when a virtual channel is estab-

lished. ATM supports switched virtual connections (SVCs) or logical links between ATM network endpoints for the duration of the connections, as well as permanent virtual connections (PVCs) that remain operational until they are no longer required (Hac, 2001).

ATM specifications facilitate implementation of a standardized infrastructure for reliable class of service (CoS) operations. ATM service classes include available bit rate (ABR), to ensure a guaranteed minimum capacity for bursty high-bandwidth traffic; constant bit rate (CBR), for fixed bit-rate transmissions of bandwidth-intensive traffic such as interactive video; and unspecified bit rate (UBR), for best-effort delivery of data-intensive traffic such as large-sized files. Also an ATM CoS, variable bit rate (VBR) defines parameters for non-real-time and real-time transmissions to ensure a specified level of throughput capacity to meet QoS requirements (Tan, Tham & Ngoh, 2003). Additionally, ATM networks define parameters for peak cell rate (PCR) and sustainable cell rate (SCR); policies for operations, administration and resource management; and protocols and mechanisms for secure transmissions (Littman, 2002).

ATM service classes combine the low delay of circuit switching with the bandwidth flexibility and high speed of packet switching. ATM switches route multiple cells concurrently to their destination, enable high aggregate throughput, and support queue scheduling and cell buffer management for realization of multiple QoS requirements (Kou, 1999). ATM employs user-to-network interfaces (UNIs) between user equipment and network switches, and network-to-network interfaces (NNIs) between network switches; and enables point-to-point, point-to-multipoint and multipoint-to-multipoint connections.

Layer 1, or the Physical Layer of the ATM protocol stack, supports utilization of diverse transmission media, interfaces and transport speeds; transformation of signals into optical/electronic formats; encapsulation of IP packets into ATM cells; and multiplexing and cell routing and switching operations. Situated above the ATM Physical Layer, Layer 2—or the ATM Layer—uses the ATM 53-byte cell as the basic transmission unit, which operates independently of the ATM Physical Layer and employs ATM switches to route cellular streams received from the ATM Adaptation Layer (AAL), or Layer 3, to destination addresses. The AAL consists of five sublayers that enable cell segmentation and re-assembly, and CBR and VBR services.

Widespread implementation of IP applications contributes to utilization of IP overlays on ATM networks. To interoperate with IP packet-switching services, ATM defines a framing structure that transports IP packets as sets of ATM cells. ATM also interworks with ISDN, frame relay, fibre channel, digital subscribe line (DSL), cable modem, GSM, UMTS and satellite technologies.

ATM SPECIFICATIONS

Standards groups in the ATM arena include the International Telecommunications Union-Telecommunications Sector (ITU-T), European Telecommunications Standards Institute (ETSI), ATM Forum and the International Engineering Task Force (IETF). Broadband passive optical networks (B-PONs) compliant with the ITU-T G.983.1 Recommendation enable optical ATM solutions that support asymmetric transmissions downstream at 622.08 Mbps (OC-12) and upstream at 155.52 Mbps (OC-3) (Effenberger, Ichibangase & Yamashita, 2001).

Sponsored by ETSI (2001), the European ATM services interoperability (EASI) and Telecommunications and IP Harmonization Over Networks (TIPHON) initiatives established a foundation for ATM interoperability operations and ATM QoS guarantees. HiperLAN2 (high performance radio local area network-2), an ETSI broadband radio access network specification, works with ATM core networks in enabling wireless Internet services at 54 Mbps.

The ATM Forum establishes ATM interworking specifications, including ATM-over-ADSL and protocols such as multi-protocol-over-ATM (MPOA) for encapsulating virtual LAN (VLAN) IP packets into ATM cells that are routed across the ATM network to destination VLAN addresses. The ATM Forum promotes integration of ATM and IP addressing schemes for enabling ATM devices to support IP version 4 (IPv4) and IPv6 operations, as well as security mechanisms and services such as elliptic curve cryptography.

Defined by the IETF, the IP multicast-over-ATM Request for Comments (RFC) supports secure delivery of IP multicasts to designated groups of multicast

recipients (Diot, Levine, Lyles, Kassem & Bolensiefen, 2000). The IETF also established RFC 2492 to support ATM-based IPv6 services. IPv6 overcomes IPv4 limitations by providing expanded addressing capabilities, a streamlined header format and merged authentication and privacy functions.

The Third Generation Partnership Project (3GPP), an international standards alliance, endorsed the use of ATM as an underlying transport technology for 3G UMTS core and access networks and satellite-UMTS (S-UMTS) configurations (Chaudhury, Mohr & Onoe, 1999). Developed by the European Space Agency (ESA) and endorsed by the ITU-T, S-UMTS supports Web browsing and content retrieval, IP multicasts and videoconferencing. S-UMTS also is a component in the suite of air interfaces for the International Mobile Telecommunications-Year 2000 (IMT-2000). This initiative enables ubiquitous mobile access to multimedia applications and communications services (Cuevas, 1999).

ATM E-LEARNING INITIATIVES

A broadband multiplexing and switching technology that supports public and private wireline and wireless operations, ATM enables tele-education applications with real-time responsiveness and high availability (Kim & Park, 2000). In this section, ATM e-learning initiatives in Estonia, Lithuania and Poland are examined. These countries also participate in EC e-learning program initiatives that support foreign language teleinstruction, intercultural exchange, pedagogical innovations in distance education and enhanced access to e-learning resources. ATM e-learning initiatives in Singapore and the U.S. are also described.

Estonia

The Estonian Education and Research Network (EENET) provisions ATM-based videoconferences and multicast distribution in the Baltic States at institutions that include the University of Tartu and Tallinn Technical University. A participant in the (networked education (NED) and Swedish-ATM (SWEST-ATM) projects, EENET also enables ATM links to the Royal Institute of Technology in Stockholm, Tampere University in Finland and the National University of Singapore (Kalja, Ots & Penjam, 1999).

Lithuania

The Lithuania Academic and Research Network (LITNET) uses an ATM backbone network to support links to academic libraries and scientific institutions; GÉANT, the pan European gigabit network; and NRENs such as EENET. The Lithuania University of Agriculture and Kaunas Medical University employ ATM and Gigabit Ethernet technologies for e-learning projects. The Kaunas Regional Distance Education Center uses ATM in concert with ISDN and satellite technologies to support access to distance education courses (Rutkauskiene, 2000).

POLAND

Sponsored by the State Committee for Scientific Research (KBN, 2000), the Polish Optical Internet (PIONIER) project employs a DWDM infrastructure that interworks with ATM, Gigabit Ethernet, IP and SDH technologies. This initiative supports e-learning applications at Polish educational institutions and research centers, including the Wroclaw University of Technology.

SINGAPORE

The Singapore Advanced Research and Education Network (SingAREN) supports ATM connections to the Asia Pacific Area Network (APAN), the Trans-Eurasia Information Network (TEIN) and the Abilene network via the Pacific Northwest GigaPoP (gigabit point of presence) in Seattle, Washington, U.S. SingAREN transborder connections enable the Singapore academic and research community to participate in global scientific investigations and advanced e-learning initiatives in fields such as space science and biology. Academic institutions in Singapore that participate in SingAREN include the National Technological University.

U.S.

Maine

The Maine Distance Learning Project (MDLP) employs ATM to support ITU-T H.323-compliant

videoconferences and facilitate access to I2 resources. The ATM infrastructure enables high school students at MDLP sites with low enrollments to participate in calculus physics and anatomy classes and take advanced college placement courses. In addition, the MDLP ATM configuration provisions links to graduate courses developed by the University of Maine faculty; certification programs for teachers, firefighters and emergency medical personnel; and teleworkshops for state and local government agencies.

New Hampshire

The Granite State Distance Learning Network (GSDLN) employs an ATM infrastructure for enabling tele-education initiatives at K-12 schools and post-secondary institutions. GSDLN provides access to professional certification programs, team teaching sessions and enrichment activities sponsored by the New Hampshire Fish and Game Department.

Rhode Island

A member of the Ocean State Higher Education Economic Development and Administrative Network (OSHEN) Consortium, an I2 special-education group participant (SEGP), Rhode Island Network (RINET) employs ATM to facilitate interactive videoconferencing and provision links to I2 e-learning initiatives. RINET also sponsors an I2 ATM virtual job-shadowing project that enables students to explore career options with mentors in fields such as surgery.

TRENDS IN ATM IMPLEMENTATION

The ATM Forum continues to support development of interworking specifications and interfaces that promote the use of ATM in concert with IP, FR, satellite and S-UMTS implementations; broadband residential access technologies such as DSL and local multipoint distribution service (LMDS), popularly called wireless cable solutions; and WDM and DWDM optical configurations. The Forum also promotes development of encapsulation methods to support converged ATM/MPLS operations for enabling ATM cells that transit IP best-effort delivery networks to provide QoS assurances. Approaches for facilitating network convergence and bandwidth consolidation by using MPLS to support an ATM overlay on an IP optical network are in development.

Distinguished by its reliable support of multimedia transmissions, ATM will continue to play a critical role in supporting e-learning applications and initiatives. In 2004, the Delivery of Advanced Network Technology to Europe (DANTE), in partnership with the Asia Europe Meeting (ASEM) initiated work on TIEN2, a South East Asia intra-regional research and education network that will support links to GÉANT (GN1), the pan-European gigabit network developed under the EC CORDIS IST initiative. GN1 employs an IP-over-SDH/WDM infrastructure that enables extremely fast transmission rates at heavily trafficked core network locations and an IP-over-ATM platform to facilitate voice, video and data transmission at outlying network sites.

The ATM Forum and the Broadband Content Delivery Forum intend to position ATM as an enabler of content delivery networks (CDNs) that deliver real-time and on-demand streaming media without depleting network resources and impairing network performance (ATM Forum, 2004). In the educational arena, ATM-based CDNs are expected to support special-event broadcasts, telecollaborative research, learner-centered instruction, Web conferencing, on-demand virtual fieldtrips and virtual training.

In addition to e-learning networks and multimedia applications, ATM remains a viable enabler of e-government, telemedicine and public safety solutions. As an example, Project MESA, an initiative developed by ETSI and the Telecommunications Industry Association (TIA), will employ a mix of ATM, satellite and wireless network technologies to support disaster relief, homeland security, law enforcement and emergency medical services (ETSI & TIA, 2004).

CONCLUSION

ATM technology is distinguished by its dependable support of e-learning applications that optimize student achievement and faculty productivity. ATM technology seamlessly enables multimedia transport, IP multicast delivery and access to content-rich Web resources with QoS guarantees. Despite technical and financial concerns, ATM remains a viable enabler of

multimedia e-learning initiatives in local and wider-area educational environments. Research and experimentation are necessary to extend and refine ATM capabilities in supporting CDNs, wireless solutions, secure network operations, and interoperability with WDM and DWDM optical networks. Ongoing assessments of ATM network performance in provisioning on-demand and real-time access to distributed e-learning applications and telecollaborative research projects in virtual environments are also recommended.

REFERENCES

ATM Forum. (2004). *Converged data networks*. Retrieved May 24, 2004, from *www.atmforum.com/downloads/CDNwhtpapr.final.pdf*

Chaudhury, P., Mohr, W., & Onoe, S. (1999). The 3GPP proposal. *IEEE Communications Magazine,* (12), 72-81.

Cuevas, E. (1999). The development of performance and availability standards for satellite ATM networks. *IEEE Communications Magazine,* (7), 74-79.

Di Concetto, M., Pavarani, G., Rosa, C., Rossi, F., Paul, S., & Di Martino, P. (1999). AMUSE: Advanced broadband services trials for residential users. *IEEE Network,* (2), 37-45.

Diot, C., Levine, B., Lyles, B. Kassem, H., & Bolensiefen, D. (2000). Deployment issues for IP multicast service and architecture. *IEEE Network,* (1), 78-88.

Effenberger, F.J., Ichibangase, H., & Yamashita, H. (2001). Advances in broadband passive optical networking technologies. *IEEE Communications Magazine,* (12), 118-124.

ETSI & TIA (2004). *Project MESA.* Retrieved June 24, 2004, from *www.projectmesa.org/home.htm*

Hac, A. (2001). Wireless ATM network architectures. *International Journal of Network Management, 11,* 161-167.

Kalja, A., Ots, A., & Penjam, J. (1999). Teleeducation projects on broadband networks in Estonia. *Baltic IT Review, 3*. Retrieved May 28, 2004, from *www.dtmedia.lv/raksti/EN/BIT/199910/99100120.stm*

KBN. (2000). *PIONIER: Polish optical Internet. Advanced applications, services and technologies for information society.* Retrieved May 22, 2004, from *www.kbn.gov/pl/en/pionier/*

Kim, W-T., & Park, Y-J. (2000). Scalable QoS-based IP multicast over label-switching wireless ATM networks. *IEEE Network,* (5), 26-31.

Kou, K. (1999). Realization of large-capacity ATM switches. *IEEE Communications Magazine,* (12), 120-133l

Lee, M.-C., Hou, C.-L., & Lee, S.-J. (2003). A simplified scheduling algorithm for cells in ATM networks for multimedia communication. *Journal of Distance Education Technologies,* (2), 37-56.

Littman, M. (2002). *Building broadband networks*. Boca Raton: CRC Press.

Parr, G., & Curran, K. (2000). A paradigm shift in the distribution of multimedia. *Communications of the ACM,* (6), 103-109.

Rutkauskiene, D. (2000). Tele-learning networks: New opportunities in the development of Lithuania's regions. *Baltic IT Review, 1*. Retrieved May 18, 2004, from *www.dtmedia.lv/raksti/en/bit/200005/00052208.stm*

Tan, S.-L., Tham, C.-K., & Ngoh, L.-H. (2003). Connection set-up and QoS monitoring in ATM networks. *International Journal of Network Management, 13,* 231-245.

TERENA. (2003). *Trans European Research and Education Association (TERENA) Compendium*. Retrieved May 18, 2004, from *www.terena.nl/compendium/2003/basicinfo.php*

KEY TERMS

10 Gigabit Ethernet: Compatible with Ethernet, Fast Ethernet and Gigabit Ethernet technologies. Defined by the IEEE 802.3ae standard, 10 Gigabit Ethernet provisions CoS or QoS assurances for multimedia transmissions, whereas ATM supports QoS guarantees.

DSL: Supports consolidation of data, video and voice traffic for enabling broadband transmissions over ordinary twisted-copper-wire telephone lines between the telephone company central office and the subscriber's residence.

E-Learning: A term used interchangeably with distance education and tele-education. E-learning refers to instructional situations in which the teacher and learner are physically separated.

H.323: An ITU-T specification that defines network protocols, operations and components for transporting real-time video, audio and data over IP networks such as the Internet and I2.

IP Multicasts: Sets of IP packets transported via point-to-multipoint connections over a network such as I2 or GÉANT to designated groups of multicast recipients. IP multicasts conserve bandwidth and network resources.

Middleware: Software that connects two or more separate applications across the Web for enabling data exchange, integration and/or support.

MPLS: Assigns a short fixed-size label to an IP packet. A streamlined version of an IP packet header, this label supports fast and dependable multimedia transmissions via label-switched paths over packet networks.

Quality of Service (QoS): Guarantees in advance a specified level of throughput capacity for multimedia transmissions via ATM networks.

SONET: Enables synchronous real-time multimedia transmission via optical fiber at rates ranging from 51.84 Mbps (OC-1) to 13.21 Gbps (OC-255). SDH is the international equivalent of SONET.

Web Cache: Stores Web content locally to improve network efficiency.

Biometric Technologies

Mayank Vatsa
Indian Institute of Technology Kanpur, India

Richa Singh
Indian Institute of Technology Kanpur, India

P. Gupta
Indian Institute of Technology Kanpur, India

A.K. Kaushik
Electronic Niketan, India

INTRODUCTION

Identity verification in computer systems is done based on measures like keys, cards, passwords, PIN and so forth. Unfortunately, these may often be forgotten, disclosed or changed. A reliable and accurate identification/verification technique may be designed using biometric technologies, which are further based on the special characteristics of the person such as face, iris, fingerprint, signature and so forth. This technique of identification is preferred over traditional passwords and PIN-based techniques for various reasons:

- The person to be identified is required to be physically present at the time of identification.
- Identification based on biometric techniques obviates the need to remember a password or carry a token.

A biometric system essentially is a pattern recognition system that makes a personal identification by determining the authenticity of a specific physiological or behavioral characteristic possessed by the user. Biometric technologies are thus defined as the "automated methods of identifying or authenticating the identity of a living person based on a physiological or behavioral characteristic." A biometric system can be either an identification system or a verification (authentication) system; both are defined below.

- **Identification:** *One to Many*—A comparison of an individual's submitted biometric sample against the entire database of biometric reference templates to determine whether it matches any of the templates.
- **Verification:** *One to One*—A comparison of two sets of biometrics to determine if they are from the same individual.

Biometric authentication requires comparing a registered or enrolled biometric sample (biometric template or identifier) against a newly captured biometric sample (for example, the one captured during a login). This is a three-step process (*Capture, Process, Enroll*) followed by a *Verification* or *Identification*.

During *Capture*, raw biometric is captured by a sensing device, such as a fingerprint scanner or video camera; then, distinguishing characteristics are extracted from the raw biometric sample and converted into a processed biometric identifier record (biometric template). Next is enrollment, in which the processed sample (a mathematical representation of the template) is stored/registered in a storage medium for comparison during authentication. In many commercial applications, only the processed biometric sample is stored. The original biometric sample cannot be reconstructed from this identifier.

BACKGROUND

Many biometric characteristics may be captured in the first phase of processing. However, automated capturing and automated comparison with previously

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

stored data requires the following properties of biometric characteristics:

- **Universal:** Everyone must have the attribute. The attribute must be one that is seldom lost to accident or disease.
- **Invariance of properties:** They should be constant over a long period of time. The attribute should not be subject to significant differences based on age or either episodic or chronic disease.
- **Measurability:** The properties should be suitable for capture without waiting time and it must be easy to gather the attribute data passively.
- **Singularity:** Each expression of the attribute must be unique to the individual. The characteristics should have sufficient unique properties to distinguish one person from any other. Height, weight, hair and eye color are unique attributes, assuming a particularly precise measure, but do not offer enough points of differentiation to be useful for more than categorizing.
- **Acceptance:** The capturing should be possible in a way acceptable to a large percentage of the population. Excluded are particularly invasive technologies; that is, technologies requiring a part of the human body to be taken or (apparently) impairing the human body.
- **Reducibility:** The captured data should be capable of being reduced to an easy-to-handle file.
- **Reliability and tamper-resistance:** The attribute should be impractical to mask or manipulate. The process should ensure high reliability and reproducibility.
- **Privacy:** The process should not violate the privacy of the person.
- **Comparable:** The attribute should be able to be reduced to a state that makes it digitally comparable to others. The less probabilistic the matching involved, the more authoritative the identification.
- **Inimitable:** The attribute must be irreproducible by other means. The less reproducible the attribute, the more likely it will be authoritative.

Among the various biometric technologies being considered are fingerprint, facial features, hand geometry, voice, iris, retina, vein patterns, palm print, DNA, keystroke dynamics, ear shape, odor, signature and so forth.

Fingerprint

Fingerprint biometric is an automated digital version of the old ink-and-paper method used for more than a century for identification, primarily by law enforcement agencies (Maltoni, 2003). The biometric device requires each user to place a finger on a plate for the print to be read. Fingerprint biometrics currently has three main application areas: large-scale Automated Finger Imaging Systems (AFIS), generally used for law enforcement purposes; fraud prevention in entitlement programs; and physical and computer access. A major advantage of finger imaging is the long-time use of fingerprints and its wide acceptance by the public and law enforcement communities as a reliable means of human recognition. Others include the need for physical contact with the optical scanner, possibility of poor-quality images due to residue on the finger such as dirt and body oils (which can build up on the glass plate), as well as eroded fingerprints from scrapes, years of heavy labor or mutilation.

Facial Recognition

Face recognition is a noninvasive process where a portion of the subject's face is photographed and the resulting image is reduced to a digital code (Zhao, 2000). Facial recognition records the spatial geometry of distinguishing features of the face. Facial recognition technologies can encounter performance problems stemming from such factors as non-cooperative behavior of the user, lighting and other environmental variables. The main disadvantages of face recognition are similar to problems of photographs. People who look alike can fool the scanners. There are many ways in which people can significantly alter their appearance, like slight change in facial hair and style.

Iris Scan

Iris scanning measures the iris pattern in the colored part of the eye, although iris color has nothing to do with the biometric[6]. Iris patterns are formed randomly. As a result, the iris patterns in the left and right eyes are different, and so are the iris patterns of identical twins. Iris templates are typically around

256 bytes. Iris scanning can be used quickly for both identification and verification applications because of its large number of degrees of freedom. Disadvantages of iris recognition include problems of user acceptance, relative expense of the system as compared to other biometric technologies and the relatively memory-intensive storage requirements.

Retinal Scan

Retinal scanning involves an electronic scan of the retina—the innermost layer of wall of the eyeball. By emitting a beam of incandescent light that bounces off the person's retina and returns to the scanner, a retinal scanning system quickly maps the eye's blood vessel pattern and records it into an easily retrievable digitized database[3]. The eye's natural reflective and absorption properties are used to map a specific portion of the retinal vascular structure. The advantages of retinal scanning are its reliance on the unique characteristics of each person's retina, as well as the fact that the retina generally remains fairly stable throughout life. Disadvantages of retinal scanning include the need for fairly close physical contact with the scanning device. Also, trauma to the eye and certain diseases can change the retinal vascular structure, and there also are concerns about public acceptance.

Voice Recognition

Voice or speaker recognition uses vocal characteristics to identify individuals using a pass-phrase (Campbell, 1997). It involves taking the acoustic signal of a person's voice and converting it to a unique digital code that can be stored in a template. Voice recognition systems are extremely well-suited for verifying user access over a telephone. Disadvantages of this biometric are that not only is a fairly large byte code required, but also, people's voices can change (for example, when they are sick or in extreme emotional states). Also, phrases can be misspoken and background noises can interfere with the system.

Signature Verification

It is an automated method of examining an individual's signature. This technology examines dynamics such as speed, direction and pressure of writing; the time that the stylus is in and out of contact with the "paper"; the total time taken to make the signature; and where the stylus is raised from and lowered onto the "paper". Signature verification templates are typically 50 to 300 bytes. The key is to differentiate between the parts of the signature that are habitual and those that vary with almost every signing. Disadvantages include problems with long-term reliability, lack of accuracy and cost.

Hand/Finger Geometry

Hand or finger geometry is an automated measurement of many dimensions of the hand and fingers. Neither of these methods takes actual prints of palm or fingers. Only the spatial geometry is examined as the user puts a hand on the sensor's surface. Hand geometry templates are typically 9 bytes, and finger geometry templates are 20 to 25 bytes. Finger geometry usually measures two or three fingers, and thus requires a small amount of computational and storage resources. The problems with this approach are that it has low discriminative power, the size of the required hardware restricts its use in some applications and hand geometry-based systems can be easily circumvented[9].

Palm Print

Palm print verification is a slightly modified form of fingerprint technology. Palm print scanning uses an optical reader very similar to that used for fingerprint scanning; however, its size is much bigger, which is a limiting factor for use in workstations or mobile devices.

Keystroke Dynamics

Keystroke dynamics is an automated method of examining an individual's keystrokes on a keyboard (Monrose, 2000). This technology examines dynamics such as speed and pressure, the total time of typing a particular password and the time that a user takes between hitting keys—dwell time (the length of time one holds down each key) as well as flight time (the time it takes to move between keys). Taken over the course of several login sessions,

these two metrics produce a measurement of rhythm unique to each user. Technology is still being developed to improve robustness and distinctiveness.

Vein Patterns

Vein geometry is based on the fact that the vein pattern is distinctive for various individuals. Vein measurement generally focuses on blood vessels on the back of the hand. The veins under the skin absorb infrared light and thus have a darker pattern on the image of the hand. An infrared light combined with a special camera captures an image of the blood vessels in the form of tree patterns. This image is then converted into data and stored in a template. Vein patterns have several advantages: First, they are large, robust internal patterns. Second, the procedure does not implicate the criminal connotations associated with the taking of fingerprints. Third, the patterns are not easily damaged due to gardening or bricklaying. However, the procedure has not yet won full mainstream acceptance. The major disadvantage of vein measurement is the lack of proven reliability[9].

DNA

DNA sampling is rather intrusive at present and requires a form of tissue, blood or other bodily sample[9]. This method of capture still has to be refined. So far, DNA analysis has not been sufficiently automatic to rank it as a biometric technology. The analysis of human DNA is now possible within 10 minutes. If the DNA can be matched automatically in real time, it may become more significant. At present, DNA is very entrenched in crime detection and will remain in the law enforcement area for the time being.

Ear Shape

Identifying individuals by ear shape is used in law enforcement applications where ear markings are found at crime scenes (Burge, 2000). Problems are faced whenever the ear is covered by hair.

Body Odor

The body odor biometrics is based on the fact that virtually every human's smell is unique. The smell is captured by sensors that are capable of obtaining the odor from non-intrusive parts of the body, such as the back of the hand. The scientific basis is that the chemical composition of odors can be identified using special sensors. Each human smell is made up of chemicals known as volatiles. They are extracted by the system and converted into a template. The use of body odor sensors broaches on the privacy issue, as the body odor carries a significant amount of sensitive personal information. It is possible to diagnose some disease or activities in last hours by analyzing body odor.

MAIN FOCUS OF THE ARTICLE

Performance Measurements

The overall performance of a system can be evaluated in terms of its *storag*e, *speed* and *accuracy*. The size of a template, especially when using smart cards for storage, can be a decisive issue during the selection of a biometric system. Iris scan is often preferred over fingerprinting for this reason. Also, the time required by the system to make an identification decision is important, especially in real-time applications such as ATM transactions.

Accuracy is critical for determining whether the system meets requirements and, in practice, the way the system responds. It is traditionally characterized by two error statistics: *False Accept Rate* (*FAR)* (sometimes called False Match Rate), the percentage of impostors accepted; and *False Reject Rate* (*FRR*), the percentage of authorized users rejected. These error rates come in pairs: For each false-reject rate there is a corresponding false alarm. In a perfect biometric system, both rates should be zero. Unfortunately, no biometric system today is flawless, so there must be a trade-off between the two rates. Usually, civilian applications try to keep both rates low. The error rate of the system when FAR equals FRR is called the *Equal Error Rate*, and is used to describe performance of the overall system. Good biometric systems have error rates of less than 1%. This should be compared to error rates in current methods of authentication, such as passwords, photo IDs, handwritten signatures and so forth. Although this is feasible in theory, practical

comparison between different biometric systems when based on different technologies is very hard to achieve. The problem with the system is that people's physical traits change over time, especially with alterations due to accident or aging. Problems can occur because of accident or aging, humidity in the air, dirt and sweat (especially with finger or hand systems) and inconsistent ways of interfacing with the system.

According to the Biometric Working Group (founded by the Biometric Consortium), the three basic types of evaluation of biometric systems are: technology, scenario and operational evaluation[9].

The goal of a *technology* evaluation is to compare competing algorithms from a single technology. The use of test sets allows the same test to be given to all participants. The goal of *scenario* testing is to determine the overall system performance in a single prototype or simulated application to determine whether a biometric technology is sufficiently mature to meet performance requirements for a class of applications. The goal of *operational* testing is to determine the performance of a complete biometric system in a specific application environment with a specific target population, to determine if the system meets the requirements of a specific application.

Problems of Using Biometric Identification

Different technologies may be appropriate for different applications, depending on perceived user profiles, the need to interface with other systems or databases, environmental conditions and a host of other application-specific parameters.

Biometrics has some drawbacks and loopholes. Some of the problems associated with biometrics systems are as follows:

- **Most of the technologies work well only for a "small" target population:** Only two biometric technologies, fingerprinting and iris scanning, have been shown in independent testing to be capable of identifying a person from a group exceeding 1,000 people. Three technologies—face, voice and signature—have been shown in independent testing to be incapable of singling out a person from a group exceeding 1,000. This can be a big problem for large-scale use[2].
- **The level of public concern about privacy and security is still high:** Privacy issues are defined as freedom from unauthorized intrusion. It can be divided into three distinct forms:
  - Physical privacy, or the freedom of individual from contact with others.
  - Informational privacy, or the freedom of individuals to limit access to certain personal information about oneself.
  - Decision privacy, or the freedom of individuals to make private choices about personal and intimate matters.

Table 1. Factors that impact any system[7]

Characteristic	Fingerprints	Face	Iris	Retina	Voice	Signature	Hand Geometry
Ease of Use	High	Medium	Medium	Low	High	High	High
Error incidence	Dryness, dirt, age	Lighting, age, glasses, hair	Poor lighting	Glasses	Noise, colds, weather	Changing signatures	Hand injury, age
Accuracy	High	High	Very high	Very high	High	High	High
User acceptance	Medium	Medium	Medium	Medium	High	Medium	Medium
Required security level	High	Medium	Very high	High	Medium	Medium	Medium
Long-term stability	High	Medium	High	High	Medium	Medium	Medium

Public resistances to these issues can be a big deterrent to widespread use of biometric-based identification.

- **Biometric technologies do not fit well in remote systems.** If verification takes place across a network (the measurement point and the access control decision point are not co-located), the system might be insecure. In this case, attackers can either steal the person's scanned characteristic and use it during other transactions or inject their characteristic into the communication channel. This problem can be overcome by the use of a secure channel between the two points.
- **Biometric systems do not handle failure well.** If someone steals one's template, it remains stolen for life. Since it is not a digital certificate or a password, you cannot ask the bank or some trusted third party to issue a new one. Once the template is stolen, it is not possible to go back to a secure situation.

CONCLUSION

The world would be a fantastic place if everything were secure and trusted. But unfortunately, in the real world there is fraud, crime, computer hackers and theft. So there is need of something to ensure users' safety. Biometrics is one method that can give optimal security to users in the available resource limitations. Some of its ongoing and future applications are:

- Physical access
- Virtual access
- E-commerce applications
- Corporate IT
- Aviation
- Banking and financial
- Healthcare
- Government

This article presents an overview of various biometrics technologies' performance, application and problems. Research is going on to provide a secure, user-friendly and cost-effective biometrics technology.

REFERENCES

Burge, M., & Burger, W. (2000). Ear biometrics for machine vision. *ICPR*, 826-830.

Campbell, J. (1997). Speaker recognition: A tutorial. *Proceedings of IEEE*, *85*(9).

Daugman, J.G. (1993). High confidence visual recognition of persons by a test of statistical independence. *IEEE PAMI*, *15*(11), 1148-1161.

Ismail, M.A., & Gad, S. (2000). Off-line Arabic signature recognition and verification. *Pattern Recognition*, *33*, 1727-1740.

Jain, A.K., Hong, L., Pankanti, S., & Bolle, R. (1997). An identity authentication system using fingerprints. *Proceedings of the IEEE*, *85*(9), 1365-1388.

Lee, L., & Grimson, W. (2002). Gait analysis for recognition and classification. *Proceedings of the International Conference on Automatic Face and Gesture Recognition.*

Maltoni, D., Maio, D., Jain, A.K., & Prabahakar, S. (2003). *Handbook of fingerprint recognition.* Springer.

Matteo, G., Dario, M., & Davide, M. (1997). On the error-reject trade-off in biometric verification systems. *IEEE PAMI*, *19*(7), 786-796.

Monrose, F., Rubin, A.D. (2000). Keystroke dynamics as a biometric for authentication. *FGCS Journal: Security on the Web*.

Nixon, M.S., Carter, J.N., Cunado, D., Huang, P.S., & Stevenage, S.V. (1999). Automatic gait recognition. *Biometrics: Personal Identification in Networked Society*, 231-249.

Zhao, W., Chellappa, R., Rosenfeld, A., & Philips, P.J. (2000). *Face recognition: A literature survey.* UMD Technical Report.

KEY TERMS

Authentication: The action of verifying information such as identity, ownership or authorization.

Biometric: A measurable, physical characteristic or personal behavioral trait used to recognize or verify the claimed identity of an enrollee.

Biometrics: The automated technique of measuring a physical characteristic or personal trait of an individual and comparing that characteristic to a comprehensive database for purposes of identification.

Behavioral Biometric: A biometric characterized by a behavioral trait learned and acquired over time.

False Acceptance Rate: The probability that a biometric system will incorrectly identify an individual or will fail to reject an impostor.

False Rejection Rate: The probability that a biometric system will fail to identify an enrollee, or verify the legitimate claimed identity of an enrollee.

Physical/Physiological Biometric: A biometric characterized by a physical characteristic.

ENDNOTES

1 http://biometrics.cse.msu.edu/

3 www.biometricgroup.com/

4 www.bioservice.ch/

6 www.biometricgroup.com/a_bio1/technology/cat_dsv.htm

7 www.computer.org/itpro/homepage/jan_feb01/security3b.htm

8 http://homepage.ntlworld.com/avanti/whitepaper.htm

9 www.biometrics.org/

Biometrics Security

Stewart T. Fleming
University of Otago, New Zealand

INTRODUCTION

Information security is concerned with the assurance of confidentiality, integrity, and availability of information in all forms. There are many tools and techniques that can support the management of information security and systems based on biometrics that have evolved to support some aspects of information security. Biometric systems support the facets of identification/authorization, authentication and non-repudiation in information security.

Biometric systems have grown in popularity as a way to provide personal identification. Personal identification is crucially important in many applications, and the upsurge in credit-card fraud and identity theft in recent years indicates that this is an issue of major concern in society. Individual passwords, PIN identification, cued keyword personal questions, or even token-based arrangements all have deficiencies that restrict their applicability in a widely-networked society. The advantage claimed by biometric systems is that they can establish an unbreakable one-on-one correspondence between an individual and a piece of data.

The drawback of biometric systems is their perceived invasiveness and the general risks that can emerge when biometric data is not properly handled. There are good practices that, when followed, can provide the excellent match between data and identity that biometrics promise; if not followed, it can lead to enormous risks to privacy for an individual.

Biometric Security

Jain et al. (2000) define a biometric security system as: …essentially a pattern-matching system which makes a personal identification by establishing the authenticity of a specific physiological or biological characteristic possessed by the user. An effective security system combines at least two of the following three elements: "something you have, something you know or something you are" (Schneier, 2000). Biometric data provides the "something you are"—data is acquired from some biological characteristic of an individual. However, biometric data is itself no guarantee of perfect security; a combination of security factors, even a combination of two or more biometric characteristics, is likely to be effective (Jain et al., 1999). Other techniques are needed to combine with biometrics to offer the characteristics of a secure system—confidentiality (privacy), integrity, authentication and non-repudiation (Clarke, 1998).

Biometric data come in several different forms that can be readily acquired, digitized, transmitted, stored, and compared in some biometric authentication device. The personal and extremely sensitive nature of biometric data implies that there are significant privacy and security risks associated with capture, storage, and use (Schneier, 1999).

Biometric data is only one component in wider systems of security. Typical phases of biometric security would include acquisition of data (the biological characteristic), extraction (of a template based on the data), comparison (with another biological characteristic), and storage. The exact design of biometric systems provides a degree of flexibility in how activities of enrollment, authentication, identification, and long-term storage are arranged. Some systems only require storage of the data locally within a biometric device; others require a distributed database that holds many individual biometric samples.

BACKGROUND

Biometric security systems can be divided logically into separate phases of operation—separating enrollment of a biometric from extraction and coding into a template form to authentication where a sample acquired from an individual at some time is compared with one enrolled at a previous time. The

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

enrollment and comparison of biometric data are done by some biometric authentication device, and a variety of biometric data can be used as the basis for the authentication. The characteristics of a number of different devices are described, and then the particular risks and issues with these devices are discussed in the main part of this article.

Types of Biometric Devices

Several types of biometric data are commonly in use. Each of the following types of devices captures data in a different form and by a different mechanism. The nature of the biometric data and the method by which they are acquired determines the invasiveness of the protocol for enrollment and authentication. The method of acquisition and any associated uncertainties in the measurement process can allow a malicious individual to attack the security of the biometric system by interfering with the capture mechanism or by substituting biometric data.

- **Fingerprint Scanner:** Acquires an image of a fingerprint either by optical scanning or capacitance sensing. Generation of biometric templates is based on matching minutiae—characteristic features in fingerprints.
- **Retinal/Iris Scanner:** Both are forms of biometric data capture based on scanning different parts of the eye. In a retinal scan, a biometric template is formed by recording the patterns of capillary blood vessels at the back of the eye. Iris scanning can be performed remotely using a high-resolution camera and templates generated by a process similar to retinal scanning.
- **Facial Scanner:** Facial recognition works by extracting key characteristics such as relative position of eyes, nose, mouth, and ears from photographs of an individual's head or face. Authentication of facial features is quite sensitive to variations in the environment (camera position, lighting, etc.) to those at enrollment.
- **Hand Geometry:** Scanners generate templates based on various features of an individual's hand, including finger length. Templates generated can be very compact, and the method is often perceived by users to be less invasive than other types of biometric devices.
- **Voiceprint:** Voiceprint recognition compares the vocal patterns of an individual with previously enrolled samples. An advantage of voiceprint techniques over other forms of biometric is the potential to detect duress or coercion through the analysis of stress patterns in the sample voiceprint.
- **DNA Fingerprint:** This method works by taking a tissue sample from an individual and then sequencing and comparing short segments of DNA. The disadvantages of the technique are in its overall invasiveness and the speed at which samples can be processed. Due to the nature of the process itself, there is an extremely low false acceptance rate, but an uncertain false rejection rate.
- **Deep Tissue Illumination:** A relatively new technique (Nixon, 2003) that involves illumination of human tissue by specific lighting conditions and the detection of deep tissue patterns based on light reflection. The technique is claimed to have less susceptibility for spoofing than other forms of biometric techniques, as it is harder to simulate the process of light reflection.
- **Keystroke Pattern:** Technique works by detecting patterns of typing on a keyboard by an individual against patterns previously enrolled. Keystroke biometrics have been used to harden password entry—to provide greater assurance that a password was typed by the same individual that enrolled it by comparing the pace at which it was typed.

Typically, the raw biometric data that are captured from the device (the measurement) are encoded into a biometric template. Extraction of features from the raw data and coding of the template are usually proprietary processes. The biometric templates are normally used as the basis for comparison during authentication. Acquisition, transmission, and storage of biometric templates are important aspects of biometric security systems, as these are areas where risks can arise and attacks on the integrity of the system can be made.

In considering the different aspects of a biometric system, we focus on the emergent issues and risks concerned with the use of this kind of data.

Careful consideration of these issues is important due to the overall concern with which users view biometric systems, the gaps between the current state of technological development, and legislation to protect the individual. In considering these issues, we present a framework based on three important principles: privacy, awareness, and control.

MAIN FOCUS

For a relatively new technology, biometric security has the potential to affect broad sectors of commerce and public society. While there are security benefits and a degree of convenience that can be offered by the use of biometric security, there are also several areas of concern. We examine here the interaction of three main issues—privacy, awareness, and consent—as regards biometric security systems, and we show how these can contribute to risks that can emerge from these systems.

Privacy

There are several aspects to privacy with relation to biometrics. First, there is the necessary invasiveness association with the acquisition of biometric data itself. Then, there are the wider issues concerned with association of such personal data with the real identity of an individual. Since biometric data can never be revoked, there are concerns about the protection of biometric data in many areas.

A biometric security system should promote the principle of authentication without identification, where possible. That is, rather than identifying an individual first and then determining the level of access that they might have, authentication without identification uses the biometric data in an anonymous fashion to determine access rights. Authentication without identification protects the privacy of the user by allowing individuals to engage in activities that require authentication without revealing their identities.

Such protection can be offered by some technologies that combine biometric authentication with encryption (Bleumer, 1998, Impagliazzo & More, 2003). However, in many situations, more general protection needs to be offered through legislation rather than from any characteristic of the technology itself. Here we find a serious gap between the state of technological and ethical or legal developments.

Legislative protections are widely variable across different jurisdictions. The United Kingdom Data Protection Act (1998), the European Union Data Protection Directive (1995), and the New Zealand Privacy Act (1994) afford protection to biometric data at the same level as personal data. In the United States, the Biometric Identifier Privacy Act in New Jersey has been enacted to provide similar levels of protection. The Online Personal Privacy Act that proposed similar protections for privacy of consumers on the Internet was introduced into the United States Senate (Hollings, 2002; SS2201 Online Personal Privacy Act, 2002) but was not completed during the session; the bill has yet to be reintroduced.

Awareness and Consent

If an individual is unaware that biometric data have been acquired, then they hardly could have given consent for it to be collected and used. Various systems have been proposed (and installed) to capture biometric data without the expressed consent of an individual, or even without informing the individual that such data is being captured. Examples of such systems include the deployment of facial recognition systems linked to crowd-scanning cameras at the Super Bowl in Tampa Bay, Florida (Wired, December 2002) or at various airports (e.g., Logan International Airport, reported in *Boston Globe*, July 2002). While it would appear from the results of such trials that these forms of biometric data acquisition/matching are not yet effective, awareness that such methods could be deployed is a major concern.

Consent presupposes awareness; however, consent is not such an easy issue to resolve with biometrics. It also presupposes that either the user has some control over how their biometric data are stored and processed, or that some suitable level of protection is afforded to the user within the context of the system. The use of strong encryption to protect biometric data during storage would be a good example of such protection. It is crucial to reach some form of agreement among all parties involved in using the system, both those responsible for authenticating and the individuals being authen-

ticated. If the user has no alternative other than to use the biometric system, can they really be said to consent to use it?

Risks

Biometric devices themselves are susceptible to a variety of attacks. Ratha, Connell & Boyle (2001) list eight possible forms of attack (Table 1) that can be used by a malicious individual to attempt to breach the integrity of a system in different ways.

Uncertainty in the precision of acquiring and comparing biometric data raises risks of different kinds associated with false acceptance and false rejection of biometric credentials. False acceptance has the more significant impact—if a user who has not enrolled biometric data is ever authenticated, this represents a serious breakdown in the security of the overall system. On the other hand, false rejection is more of an inconvenience for the individual—they have correctly enrolled data, but the device has not authenticated them for some reason. The degree of uncertainty varies between devices for the same type of biometric data and between different types of biometrics. Adjusting the degree of uncertainty of measurement allows the designer of a biometric security system to make the appropriate tradeoffs between security and convenience.

Biometrics are not secrets (Schneier, 1999). If biometric data are ever compromised, it raises a significant problem for an individual. If the data are substituted by a malicious individual, then the future transactions involving their credentials are suspect. Biometric data can never be revoked and, hence, should be afforded the highest protection. Fingerprint-based biometrics, for example, are relatively commonly used, and yet fingerprints are easily compromised and can even be stolen without the knowledge of the individual concerned.

The class of attacks noted as spoofing exploit this uncertainty and allow the integrity of a biometric system to be undermined by allowing fake biometric data to be introduced. We examine next how this class of attack can be conducted.

SPOOFING BIOMETRIC SECURITY

Spoofing is a class of attack on a biometric security system where a malicious individual attempts to circumvent the correspondence between the biometric data acquired from an individual and the individual itself. That is, the malicious individual tries to introduce fake biometric data into a system that does not belong to that individual, either at enrollment and/or authentication.

The exact techniques for spoofing vary, depending on the particular type of biometric involved. Typically though, such methods involve the use of some form of prosthetic, such as a fake finger, substitution of a high-resolution image of an iris, a mask, and so forth. The degree of veracity of the prosthetic varies according to the precision of the biometric device being spoofed and the freedom that the attacker has in interacting with the device. It is surprising how relatively simple methods can be successful at circumventing the security of commonly available contemporary biometric devices

Table 1. Types of attack on a biometric system

• Generic attacks
• Presentation of a fake biometric (spoofing)
• Replay attack (pre-recorded biometric data
• Interference with biometric feature extraction
• Interference with template generation
• Data substitution of biometric in storage
• Interception of biometric data between device and storage
• Overriding the final decision to match the biometric data
• Specific attacks
• Dummy silicone fingers, duplication with and without cooperation (van der Putte and Keuning, 2000)
• Present a fake fingerprint based on a gelatine mould (Matsumoto, 2002)
• Present fake biometrics or confuse the biometric scanners for fingerprints, facial recognition and retinal scanners (Thalheim et al., 2002)

(Matsumoto, 2002; Thalheim et al., 2002). Reducing the freedom that a potential attacker has via close supervision of interaction with the authentication device may be a solution; incorporation of different security elements into a system is another.

Two- or even three-factor (inclusion of two or three of the elements of security from Schneier's definition) security systems are harder to spoof; hence, the current interest in smart-cards and embedded authentication systems where biometric authentication is integrated with a device that the individual carries and uses during enrollment and authentication. A wider solution is the notion of a competitive or adversarial approach to verifying manufacturers' claims and attempting to circumvent biometric security (Matsumoto, 2002). Taking the claims made by manufacturers regarding false acceptance and false rejection rates and the degree to which their products can guarantee consideration only of live biometric sources is risky and can lead to a reduction in overall system integrity.

CONCLUSION

While biometric security systems can offer a high degree of security, they are far from perfect solutions. Sound principles of system engineering are still required to ensure a high level of security rather than the assurance of security coming simply from the inclusion of biometrics in some form.

The risks of compromise of distributed database of biometrics used in security applications are high, particularly where the privacy of individuals and, hence, non-repudiation and irrevocability are concerned (see Meeks [2001] for a particularly nasty example). It is possible to remove the need for such distributed databases through the careful application of biometric infrastructure without compromising security.

The influence of biometric technology on society and the potential risks to privacy and threats to identity will require mediation through legislation. For much of the short history of biometrics, the technological developments have been in advance of the ethical or legal ones. Careful consideration of the importance of biometric data and how they should be legally protected is now required on a wider scale.

REFERENCES

Clarke, R. (1998). Cryptography in plain text. *Privacy Law and Policy Reporter, 3*(2), 24-27.

Hollings, F. (2002). *Hollings introduces comprehensive online privacy legislation*. Retrieved from *http://hollings.senate.gov/~hollings/press/2002613911.html*

Jain, A., Hong, L., & Pankanti, S. (2000). Biometrics: Promising frontiers for emerging identification market. *Communications of the ACM, 43*(2), 91-98.

Jain, A.K., Prabhakar, S., & Pankanti, S. (1999). Can multi-biometrics improve performance? *Proceedings of the AutoID '99*, Summit, NJ.

Matsumoto, T. (2002). *Gummy and conductive silicone rubber fingers: Importance of vulnerability analysis.* In Y. Zheng (Ed.), *Advances in cryptology—ASIACRYPT 2002* (pp. 574-575). Queenstown, New Zealand.

Meeks, B.N. (2001). Blanking on rebellion: Where the future is "Nabster." *Communications of the ACM, 44*(11), 17.

Nixon, K. (2003). Research & development in biometric anti-spoofing. *Proceedings of the Biometric Consortium Conference*, Arlington, VA.

Ratha, N.K., Cornell, J.H., & Bolle, R.M. (2001). A biometrics-based secure authentication system. *IBM Systems Journal, 40*(3), 614-634.

S2201 online personal privacy act: Hearing before the Committee on Commerce, Science and Transportation. (2002). *United States Senate, 107th Sess*.

Schneier, B. (1999). Biometrics: Uses and abuses. *Communications of the ACM, 42*(8), 136.

Schneier, B. (2000). *Secrets and lies: Digital security in a networked world*. New York: Wiley.

Thalheim, L., Krissler, J., & Ziegler, P.-M. (2002, November). Body check—Biometric access protection devices and their programs put to the test. *c't Magazine*, 114.

Tomko, G. (1998). Biometrics as a privacy-enhancing technology: Friend or foe of privacy. *Proceed-*

ings of the Privacy Laws and Business Privacy Commissioners / Data Protection Authorities Workshop, Santiago de Compostela, Spain.

van der Putte, T., & Keuning, J. (2000). Biometrical fingerprint recognition: Don't get your fingers burned. *Proceedings of the Fourth Working Conference on Smart Card Research and Advanced Applications*, Bristol, UK.

KEY TERMS

Authentication: The process by which a contemporary biometric sample is acquired from an individual and used to compare against a historically enrolled sample. If the samples match, the user is authenticated. Depending on the type of system, the authentication may be prompted by some additional information—a key to the identity of the user or the pseudonym against which the enrolled data was registered.

Biometric: Some measurement of the biological characteristics of a human subject. A useful biometric is one that is easily acquired and digitized and where historical samples can be readily compared with contemporary ones.

Biometric Encryption: A technique whereby the biometric data is used as a personal or private key to be used in some cryptographic process.

Enrollment: The initial acquisition and registration of biometric data for an individual. Dependent on the type of biometric system, this data may be registered in association with the identity of the user or against some pseudonym that preserves anonymity.

False Acceptance: A case where an individual is authenticated when they were not the person that enrolled the original sample.

False Rejection: A case where an individual is not authenticated, although they have previously enrolled biometric data.

Irrevocability: The inability of an individual to be able to somehow cancel some credential. Biometric systems run a high risk of compromising irrevocability, if biometric data belonging to an individual is ever acquired and used to spoof a system.

Non-Repudiation: The inability of an individual to disavow some action or his or her presence at a particular location at some specific time. Biometric security systems have the potential to offer a high degree of non-repudiation due to the intimately personal nature of biometric data.

Spoofing: An activity where a malicious individual aims to compromise the security of a biometric system by substituting fake biometric data in some form or another. Anti-spoofing techniques are measures designed to counteract spoofing activities.

Biometrics, A Critical Consideration in Information Security Management

Paul Benjamin Lowry
Brigham Young University, USA

Jackson Stephens
Brigham Young University, USA

Aaron Moyes
Brigham Young University, USA

Sean Wilson
Brigham Young University, USA

Mark Mitchell
Brigham Young University, USA

INTRODUCTION

The need for increased security management in organizations has never been greater. With increasing globalization and the spread of the Internet, information-technology (IT) related risks have multiplied, including identity theft, fraudulent transactions, privacy violations, lack of authentication, redirection and spoofing, data sniffing and interception, false identities, and fraud.

Many of the above problems in e-commerce can be mitigated or prevented by implementing controls that improve authentication, nonrepudiation, confidentiality, privacy protection, and data integrity (Torkzadeh & Dhillon, 2002). Several technologies help support these controls, including data encryption, trusted third-party digital certificates, and confirmation services. Biometrics is an emerging family of authentication technologies that supports these areas.

It can be argued that authentication is the baseline control for all other controls; it is critical in conducting e-commerce to positively confirm that the people involved in transactions are who they say they are. Authentication uses one or more of the following methods of identification (Hopkins, 1999): something you know (e.g., a password), something you have (e.g., a token), and something about you (e.g., a fingerprint). Using knowledge is the traditional approach to authentication, but it is the most prone to problems, because this knowledge can be readily stolen, guessed, or discovered through computational techniques. Physical objects tend to be more reliable sources of identification, but this approach suffers from the increased likelihood of theft. The last approach to authentication is the basis for biometrics. Biometrics refers to the use of computational methods to evaluate the unique biological and behavioral traits of people (Hopkins, 1999) and it is arguably the most promising form of authentication because personal traits (e.g., fingerprints, voice patterns, or DNA) are difficult to steal or emulate.

BACKGROUND

A given biometric can be based on either a person's physical or behavioral characteristics. Physical characteristics that can be used for biometrics include fingerprints, hand geometry, retina and iris patterns, facial characteristics, vein geometry, and DNA. Behavioral biometrics analyze how people perform actions, including voice, signatures, and typing patterns.

Biometrics generally adhere to the following pattern: When a person first "enrolls" in a system, the

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

target biometric is scanned and stored as a template in a database that represents the digital form of the biometric. During subsequent uses of the system the biometric is scanned and compared against the stored template.

The process of scanning and matching can occur through verification or identification. In verification (a.k.a. authentication) a one-to-one match takes place in which the user must claim an identity, and the biometric is then scanned and checked against the database. In identification (a.k.a. recognition), a user is not compelled to claim an identity; instead, the biometric is scanned and then matched against all the templates in the database. If a match is found, the person has been "identified."

The universal nature of biometrics enables them to be used for verification and identification in forensic, civilian, and commercial settings (Hong, Jain, & Pankanti, 2000). Forensic applications include criminal investigation, corpse identification, and parenthood determination. Civilian uses include national IDs, driver's licenses, welfare disbursement, national security, and terrorism prevention. Commercial application includes controlling access to ATMs, credit cards, cell phones, bank accounts, homes, PDAs, cars, and data centers.

Despite the promise of biometrics, their implementation has yet to become widespread. Only $127 million was spent on biometric devices in the year 2000, with nearly half being spent on fingerprinting; however, future growth is expected to be strong, with $1.8 billion worth of biometrics-related sales predicted in 2004 (Mearian, 2002). Clearly, the true potential of biometrics has yet to be reached, which opens up many exciting business and research opportunities. The next section reviews specific biometrics technologies.

BIOMETRICS TECHNOLOGIES

This section reviews the major biometrics technologies and discusses where they are most appropriate for use. We examine iris and retina scanning, fingerprint and hand scanning, facial recognition, and voice recognition.

Retina and Iris Scanning

Considered by many to be the most secure of all biometrics, eye-based biometrics have traditionally been utilized in high-security applications, such as prisons, government agencies, and schools. Eye scanning comes in two forms: iris scanning and retina scanning. The first biometric eye-scanning technologies were developed for retina recognition. Retinal scanners examine the patterns of blood vessels at the back of the eye by casting either natural or infrared light onto them. Retina scanning has been demonstrated to be an extremely accurate process that is difficult to deceive because retinal patterns are stable over time and unique to individuals (Hong et al., 2000).

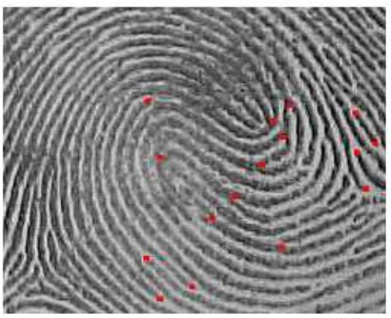
Iris scanning is a newer technology than retina scanning. The iris consists of the multicolored portion of the eye that encircles the pupil, as shown in Figure 1. Iris patterns are complex, containing more raw information than a fingerprint. The iris completes development during a person's first two years of life, and its appearance remains stable over long periods of time. Irises are so personally unique that even identical twins exhibit differing iris patterns.

Two differences between retina and iris scanning are the equipment and the procedures. The equipment for retina recognition tends to be bulky and complex and the procedures tend to be uncomfortable. Users must focus on a particular spot for a few seconds and their eyes must be up close to the imaging device. Figure 2 shows an iris scanner sold by Panasonic. Unlike retinal scanning, iris recognition involves more standard imaging cameras that are not as specialized or as expensive. Iris scanning can be accomplished

Figure 1. Depiction of an iris from www.astsecurity.com

Figure 2. Panasonic BM-ET100US authenticam iris recognition camera

Figure 3. Depiction of fingerprint ridges from www.windhampolice.com



with users situated at a distance of up to one meter away from the camera. Another difference is that retinal scans require people to remove their glasses, whereas iris scans work with glasses. Iris scanners also detect artificial irises and contact lenses.

In terms of accuracy, retina scanning has a proven track record; hence, it is used more in high-security installations. Because iris systems are newer they have less of a track record. Although template-matching rates are fairly high for both technologies, preliminary results indicate that iris recognition excels at rejecting unauthorized users but also frequently denies authorized user (false negatives).

Compared to other biometrics devices, eye-scanning equipment is expensive. Retinal imaging is especially costly because the required equipment is similar to specialized medical equipment, such as a retinascope, whereas iris recognition uses more standard and inexpensive cameras.

Fingerprint Scanning

Fingerprint scanning uses specialized devices to capture information about a person's fingerprint, which information is used to authenticate the person at a later time. Each finger consists of unique patterns of lines. Fingerprint scanners do not capture entire fingerprints; instead, they record small details about fingerprints, called minutiae (Hong et al., 2000). For example, a scanner will pick a point on a fingerprint and record what the ridge at that point looks like (as seen in Figure 3), which direction the ridge is heading, and so on (Jain, Pankanti, & Prabhakar, 2002). By picking enough points, the scanner can be highly accurate. Although minutiae identification is not the only suitable factor for fingerprint comparison, it is the primary feature used by fingerprint systems. The number of minutiae per fingerprint can vary, but a high-quality fingerprint scan will contain between 60 and 80 minutiae (Hong et al., 2000).

A biometrics system can identify a fingerprint from its ridge-flow pattern; ridge frequency; location and position of singular points; type, direction, and location of key points; ridge counts between pairs of minutiae; and location of pores (Jain et al., 2002). Given their simplicity and multiple uses, fingerprint scanning is the most widely used biometrics application.

One significant point is that vulnerabilities abound throughout the entire process of fingerprint authentication. These vulnerabilities range from the actual scan of the finger to the transmission of the authentication request to the storing of the fingerprint data. Through relatively simple means, an unauthorized person can gain access to a fingerprint-scanning system (Thalheim, Krissler, & Ziegler, 2002): the scanners may be deceived by simply blowing on the scanner surface, rolling a bag of warm water over it, or using artificial wax fingers. Another weakness with some fingerprint scanners is the storage and transmission of the fingerprint information. Fingerprint minutiae are stored as templates in databases on servers; thus, the inherent vulnerability of a computer network becomes a weakness. The fingerprint data must be transmitted to the server, and the transmission process may not be secure. Additionally, the fingerprint templates on a server must be protected by firewalls, encryption, and other basic network security measures to keep the templates secure.

An organization's size is another critical component in determining the effectiveness of a fingerprint system. Larger organizations require more time and resources to compare fingerprints. Although this is not an issue for many organizations, it can be an issue for large and complex government organizations such as the FBI (Jain et al., 2002).

Variances in scanning can also be problematic because spurious minutiae may appear and genuine minutiae may be left out of a scan, thus increasing the difficulty of comparing two different scans (Kuosmanen & Tico, 2003). Each scan of the same fingerprint results in a slightly different representation. This variance is caused by several factors, including the position of the finger during the scan and the pressure of the finger being placed on the scanner.

Facial Recognition

One of the major advantages of facial recognition over other biometric technologies is that it is fairly nonintrusive. Facial recognition does not require customers to provide fingerprints, talk into phones, nor have their eyes scanned. As opposed to hand-based technologies, such as fingerprint scanners, weather conditions and cleanliness do not strongly affect the outcome of facial scans, making facial recognition easier to implement.

However, more than other physical biometrics, facial recognition is affected by time. The appearance and shape of a face change with one's aging process and alterations to a face—through surgery, accidents, shaving, or burns, for example—can also have a significant effect on the result of facial-recognition technology.

Thus far, several methods of facial recognition have been devised. One prominent technique analyzes the bone structure around the eyes, nose, and cheeks. This approach, however, has several limitations. First, the task of recognizing a face based on images taken from different angles is extremely difficult. Furthermore, in many cases the background behind the subject must be overly simple and not representative of reality (Hong et al., 2000).

Technology also exists that recognizes a neural-network pattern in a face and scans for "hot spots" using infrared technology. The infrared light creates a so-called "facial thermogram" that overcomes some of the limitations normally imposed on facial recognition. Amazingly, plastic surgery that does not alter the blood flow beneath the skin and rarely affects facial thermograms (Hong et al., 2000). A facial thermogram can also be captured in poorly lit environments. However, research has not yet determined if facial thermograms are adequately discriminative; for example, they may depend heavily on the emotion or body temperature of an individual at the moment the scan is created (Hong et al., 2000).

A clear downside to facial recognition is that it can more easily violate privacy through powerful surveillance systems. Another problem specific to most forms of facial recognition is the requirement of bright lights and a simple background. Poor lighting or a complex background can make it difficult to obtain a correct scan. Beards and facial alterations can also negatively affect the recognition process.

Voice Recognition

Voice recognition differs from most other biometric models in that it uses acoustic information instead of images. Each individual has a unique set of voice characteristics that are difficult to imitate. Human speech varies based on physiological features such as the size and shape of an individual's lips, nasal cavity, vocal chords, and mouth (Hong et al., 2000).

Voice recognition has an advantage over other biometrics in that voice data can be transmitted over phone lines, a feature that lends to its widespread use in such areas as security, fraud prevention, and monitoring (Markowitz, 2000). Voice recognition has shown success rates as high as 97%. Much of this success can be explained by the way a voice is analyzed when sample speech is requested for validation.

Voice biometrics use three types of speaker verification: text dependent, text prompted, and text independent. Text-dependent verification compares a prompted phrase, such as an account number or a spoken name, to a prerecorded copy of that phrase stored in a database. This form of verification is frequently used in such applications as voice-activated dialing in cell phones and bank transactions conducted over a phone system.

Text-prompted verification provides the best alternative for high-risk systems. In this case, a sys-

tem requests multiple random phrases from a user to lessen the risk of tape-recorded fraud. The main drawback to this verification process is the amount of time and space needed to create a new user on the system (Markowitz, 2000). This procedure is often used to monitor felons who are under home surveillance or in community-release programs.

Text-independent verification is the most difficult of the three types of voice recognition because nothing is asked of the user. Anything spoken by the user can be used to verify authenticity, a process which can make the authentication process virtually invisible to the user.

One drawback of voice recognition technique is that it is increasingly difficult to manage feedback and other forms of interference when validating a voice. Voices are made up entirely of sound waves. When transmitted over analog phone lines these waves tend to become distorted. Current technologies can reduce noise and feedback, but these problems cannot be entirely eliminated.

Voice-recognition products are also limited in their ability to interpret wide variations of voice patterns. Typically, something used for purposes of authentication must be spoken at a steady pace without much enunciation or pauses. Yet human speech varies so greatly among individuals that it is a challenge to design a system that will account for variations in speed of speech as well as in enunciation.

Despite its imperfections, voice recognition has a success rate of up to 98%. Consequently, whereas about 2% of users will be declined access when they are indeed who they say they are, only about 2% of users will be granted access when they are not who they say they are.

PRACTITIONER IMPLICATIONS

To help practitioners compare these biometrics, we present Table 1 to aid with decisions in implementing biometrics. This table compares the five major areas of biometrics based on budget consciousness, ease of use, uniqueness, difficulty of circumvention, space savings, constancy over time, accuracy, and acceptability by users. Each area is rated as follows: VL (very low), L (low), M (medium), H (high), and VH (very high).

FUTURE TRENDS

One area in biometrics in which much work still needs to be done is receiver operating characteristics (ROC). ROC deals with system accuracy in certain environments, especially as it relates to false-positive and false-negative results. False posi-

Table 1. Comparing biometrics

BIOMETRICS RELATIVE COMPARISON MATRIX					
	Retina Scanning	Iris Scanning	Fingerprint Scanning	Facial Recognition	Voice Recognition
Budget Consciousness	VL	L	H	M	VH
Ease of Use	VL	L	M	VH	H
Uniqueness of Biometric	H	VH	M	L	VL
Difficulty of Circumvention	VH	H	M	L	VL
Space Savings	VL	L	H	M	VH
Constancy over Time	H	VH	M	L	VL
Accuracy	VH	H	M	VL	L
Acceptability by Users	VL	L	M	VH	H

tives, also known as false match rates (FMR), occur when an unauthorized user is authenticated to a system. False negatives, also known as false nonmatch rates (FNR), occur when an authorized user is denied access to a system. Both situations are undesirable. Unfortunately, by making one less likely, the other becomes more likely. This difficult tradeoff can be minimized by achieving a proper balance between the two extremes of strictness and flexibility. To this end, most biometrics implementations incorporate settings to adjust the degree of tolerance. In general, more secure installations require a higher degree of similarity for matches to occur than do less secure installations.

Research should also be undertaken to address three areas of attack to which biometrics are most susceptible: (1) copied-biometric attacks, in which obtaining a substitute for a true biometric causes proper authentication to occur via the normal system procedures; (2) replay attacks, in which perpetrators capture valid templates and then replay them to biometrics systems; (3) and database attacks, in which perpetrators access a template database and obtain the ability to replace valid templates with invalid ones.

Cancelable biometrics may reduce the threat of these attacks by storing templates as distortions of biometrics instead of the actual biometrics themselves (Bolle, Connell, & N., 2001). Similar to how a hash function works, the actual biometrics are not recoverable from the distortions alone. When a user is first enrolled in a system the relevant biometric is scanned, a distortion algorithm is applied to it, and the distortion template is created. Thereafter, when a scan of the biometric is taken, it is fed through the algorithm to check for a match.

Other possibilities for reducing attacks on biometrics include using biometrics that are more difficult to substitute, including finger length, wrist veins (underside), hand veins (back of hand), knuckle creases (while gripping something), fingertip structure (blood vessels), finger-section lengths, ear shape, lip shape, brain scans, and DNA (Smith, 2003). DNA is particularly intriguing because it is universal and perfectly unique to individuals.

CONCLUSION

A single biometrics system alone likely is not an ideal form of security, just as a lone username-password pair is rarely desirable for secure installations. Instead, we recommend that biometrics be implemented in combinations. This can be accomplished through multifactor authentication that mixes something you know with something you have and something about you, or through hybrid-biometrics systems that take advantage of more than one biometric to achieve better results.

As we have demonstrated, none of the most commonly used biometrics are without flaws. Some are very expensive, others are difficult to implement, and some are less accurate. Yet biometrics hold a bright future. This emerging family of technologies has the capability of improving the lives of everyone as they become a standard part of increasing the security of everyday transactions, ranging from ATM withdrawals to computer log-ins. Well-intentioned and well-directed research will help further the effective widespread adoption of biometric technologies.

REFERENCES

Bolle, R., Connell, J., & N, R. (2001). Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal, 40*(3), 628-629.

Hong, L., Jain, A. & Pankanti, S. (2000). Biometric identification. *Communications of the ACM (CACM), 43*(2), 91-98.

Hong, L., Pankanti, S. & Prabhakar, S. (2000). Filterbank-based fingerprint matching. *IEEE Transactions on Image Processing, 9*(5), 846-859.

Hopkins, R. (1999). An introduction to biometrics and large scale civilian identification. *International Review of Law, Computers & Technology, 13*(3), 337-363.

Jain, A., Pankanti, S., & Prabhakar, S. (2002). On the individuality of fingerprints. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(8), 1010-1025.

Kuosmanen, P. & Tico, M. (2003). Fingerprint matching using an orientation-based minutia descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(8), 1009-1014.

Markowitz, J. (2000). Voice biometrics. *Communications of the ACM (CACM), 43*(9), 66-73.

Mearian, L. (2002). Banks eye biometrics to deter consumer fraud. *Computerworld, 36*(5), 12.

Smith, C. (2003). The science of biometric identification. *Australian Science Teachers' Journal, 49*(3), 34-39.

Thalheim, L., Krissler, J., & Ziegler, P.-M. (2002). *Body check: Biometrics defeated*. Retrieved May 04, 2004, from *http://www.extremetech.com/article2/0%2C1558%2C13919%2C00.asp*

Torkzadeh, G. & Dhillon, G. (2002). Measuring factors that influence the success of Internet commerce. *Information Systems Research (ISR), 13*(2), 187-206.

KEY TERMS

Authentication: Guarantees that an individual or organization involved in a transaction are who they say they are.

Biometrics: The use of computational methods to evaluate the unique biological and behavioral traits of people.

Confidentiality: Guarantees that shared information between parties is only seen by authorized people.

Data Integrity: Guarantees that data in transmissions is not created, intercepted, modified, or deleted illicitly.

Identification: A user is not compelled to claim an identity first; instead, the biometric is scanned and then matched against all the templates in the database (also referred to as recognition).

Nonrepudiation: Guarantees that participants in a transaction cannot deny that they participated in the transaction.

Privacy Protection: Guarantees that shared personal information will not be shared with other parties without prior approval.

Verification: A one-to-one match with a biometric takes place during which the user must claim an identity first and then is checked against their identity (also referred to as authentication).

Broadband Solutions for Residential Customers

Mariana Hentea
Southwestern Oklahoma State University, USA

HOME NETWORKING

The term "home networking" implies that electronic network devices work together and communicate amongst themselves. These devices are classified in three categories: appliances, electronics and computers. Home networks include home theater, home office, small office home office (SOHO), intelligent appliances, smart objects, telecommunications products and services, home controls for security, heating/cooling, lighting and so forth. The suite of applications on each device, including the number of connected devices, is specific to each home. The home network configurations are challenges, besides the unpredictable problems that could be higher compared to a traditional business environment. These are important issues that have to be considered by developers supporting home networking infrastructure. In addition, home networks have to operate in an automatically configured plug-and-play mode. Home networks support a diverse suite of applications and services discussed next.

BROADBAND APPLICATIONS

Home networks carry phone conversations, TV programs and MP3 music programs, link computers and peripherals, electronic mail (e-mail), distribute data and entertainment programs, Internet access, remote interactive services and control of home appliances, lights, temperature and so forth. The most important remote interactive services include remote metering, home shopping, medical support, financial transactions, interactive TV, video telephony, online games, voice-over Internet Protocol (VoIP) and so forth. Home applications based on multimedia require Internet connections and higher data transfer rates. For example, video programs compressed to MPEG-2 standards require a 2-4 Mbps transfer rate; DVD video requires 3-8 Mbps; and high-definition TV requires 19 Mbps. Since the existing phone line connected to a modem does not support data rates higher than 56 Kbps, rather than installing a modem for each computer, the high-speed connection may be provided by a single access point called broadband access. Broadband access provides information and communication services to end users with high-bandwidth capabilities. The next section provides an overview of broadband access solutions.

BROADBAND ACCESS SOLUTIONS

The circuit between a business or home and the local telephone company's end office is called a local loop. Originally, local-loop service carried only telephone service to subscribers. But today, several local-loop connection options are available from carriers. These include dial-up circuits, Integrated Services Digital Network (ISDN) and broadband. "Last mile" refers to the telecommunication technology that connects a subscriber's home directly to the cable or telephone company. Broadband transmission is a form of data transmission in which a single medium can carry several channels at once. The carrying capacity medium is divided into a number of subchannels; each subchannel transports traffic such as video, low-speed data, high-speed data and voice (Stamper & Case, 2003). The broadband access options include Digital Subscriber Line (DSL), cable modems, broadband integrated services digital network (B-ISDN) line, broadband power line and broadband wireless, with a data rate varying from hundreds of Kbps to tens of Mbps.

Digital Subscriber Line

DSL is a technique for transferring data over regular phone lines by using a frequency different from

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

traditional voice calls or analog modem traffic over the phone wires. DSL requires connection to a central telephone office, usually less than 20,000 feet. DSL lines carry voice, video and data, and DSL service provides transmission rates to maximum 55 Mbps, which is faster than analog modems and ISDN networks. In addition to high-speed Internet access, DSL provides other services, such as second telephone line on the same pair of wires, specific broadband services, video and audio on demand. The priority between these services depends on the users and geographical area. For example, Asian users demand video services, while North American telephone companies use it for Internet access service.

Globally, the DSL market reached 63.8 million subscribers by March 2003, and future growth is expected to reach 200 million subscribers—almost 20% of all phone lines—by the end of 2005 (DSL Forum Report, 2003). xDSL refers to all types of DSL technologies, classified into two main categories: symmetric (upstream and downstream data rates are equal) and asymmetric (upstream and downstream data rates are different). DSL services include asymmetric DSL (ADSL), rate-adaptive DSL (RADSL), high data-rate DSL (HDSL), symmetric DSL (SDSL), symmetric high data-rate DSL (SHDSL) and very high data-rate DSL (VDSL) with data rates scaling with the distance and specific to each technology. For example, ADSL technology supports downstream data rates from 1.5 Mbps to 9 Mbps and upstream data rates up to 1 Mbps. VDSL technology supports downstream rates up to 55 Mbps. Also, VDSL provides bandwidth performance equal to the optical fiber, but only over distances less than 1,500 meters. SDSL technology provides data rates up to 3 Mbps. SHDSL supports adaptive symmetrical data rates from 192 Kbps to 2.31 Mbps with increments of 8 Kbps on single pair of wire or 384 Kbps to 4.6 Mbps with increments of 16 Kbps on dual pair of wire. SDSL has been developed as a proprietary protocol in North America, but it is now moving to an international standard called G.SHDSL or G.991.2. This is the first technology developed as an international standard by the International Telecommunications Union (ITU). It incorporates features of other DSL technologies and transports T1, E1, ISDN, ATM and IP signals. ADSL service is more popular in North America, whereas SDSL service is being used as a generic term in Europe to describe the G.SHDSL standard of February 2001.

Cable Access

Cable access is a form of broadband access using a cable modem attached to a cable TV line to transfer data with maximum downstream rates of 40 Mbps and upstream rates of 320 Kbps to 10 Mbps. Cable services include Internet access, telephony, interactive media, video on demand and distance learning. Networks built using Hybrid Fiber-Coax (HFC) technologies can transmit analog and digital services simultaneously. The central office transmits signals to fiber nodes via fiber-optic cables and feeders. The fiber node distributes the signals over coaxial cable, amplifiers and taps out to business users and customer service areas that consist of 500 to 2,000 home networks with a data rate up to 40 Mbps. Cable companies gained lots of users in the United States (U.S.) and expect 24.3 million cable modems installed by the end of 2004, which represents an increase from 1.2 million cable modems installed in 1998. Cable services are limited by head-end and fiber-optic installation. HFC is one possible implementation of a passive optical network (PON). Fiber to the curb can provide higher bit rates; roughly 40 times the typical rates with a cable modem (Cherry, 2003).

Fiber-optic cable is used by telephone companies in place of long-distance wires and increasingly by private companies in implementing local data communication networks. Although the time for the massive introduction of fiber is quite uncertain, the perseverance of the idea of fiber in the loop (FITL) lies in the fact that the costs of optics are coming down, bandwidth demand is going up and optical networking spreads in metropolitan areas. Because the data over cable travels on a shared loop, customers see data transfer rates drop as more users gain service access.

Broadband Wireless Access

Wire line solutions did not secure telecommunication operators, because costs and returns on investments are not scalable with the number of attached users. Although various broadband access solutions (like DSL, cable, FITL) were implemented, the killer application video on demand disappeared for the benefit of less-demanding Web access. Unsatisfactory progress of wire line solutions pushed alternative solutions based on wireless technologies. Broadband wireless access (BWA) has emerged as a technology,

which is profitable. Broadband wireless access is part of wireless local loop (WLL), radio local loop (RLL) and fixed wireless access (FWA).

WLL systems are based on a range of radio technologies such as satellite, cellular/cordless and many narrowband and broadband technologies. One WLL approach is placing an antenna on a utility pole (or another structure) in a neighborhood. Each antenna is capable of serving up to 2,000 homes. Subscribers must have an 18-inch antenna installed on their homes.

RLL systems connect mobile terminals at least in highly crowded areas to the point of presence of the operator's cable-based Asynchronous Transfer Mode (ATM) backbone network.

FWA systems support wireless high-speed Internet access and voice services to fixed or mobile residential customers located within the reach of an access point or base transceiver station. FWA systems promise rapid development, high scalability and low maintenance.

TRENDS

The two emerging broadband access technologies include fiber access optimized for clusters of business customers and Wireless LAN (WLAN) to provide service to small business and home subscribers. Use of wireless, DSL and cable for broadband access has become increasingly prevalent in metropolitan areas.

Vast geographic regions exist where broadband services are either prohibitively expensive or simply unavailable at any price. Several alternatives are emerging for using 2.4 GHz band specified in IEEE 802.11b and IEEE 802.11g protocols. The use of 5 GHz band is specified in IEEE 802.11a protocol. IEEE 802.11a and IEEE 802.11b operate using radio frequency (RF) technology and together are called Wireless Fidelity (WiFi) technology. However, WiFi technology based on IEEE 202.11b is used more for home networks. WiFi opens new possibilities for broadband fixed wireless access. There are differences on capabilities supported by these specifications. Public use of WiFi is emerging in hot spots deployed in hotels, airports, coffee shops and other public places. Hot spots are expanded to hot zones that cover a block of streets. WiFi-based broadband Internet access is also financially viable in a rural area, because it can provide fixed broadband access for towns, smaller remote communities, clusters of subscribers separated by large intercluster distances, as well as widely scattered users (Zhang & Wolff, 2004). The companies typically utilize WiFi for last mile access and some form of radio link for backhaul, as well. The proliferation of WiFi technology resulted in significant reductions in equipment costs, with the majority of new laptops now being shipped with WiFi adapters built in. The network consists of wireless access points serving end users in a point-to-multipoint configuration, interconnected to switches or routers using point-to-point wireless backhaul.

Both broadband wireless access and mobile multimedia services are a challenge for the research in wireless communication systems, and a new framework, Multiple-Input Multiple-Output (MIMO), is proposed (Murch & Letaief, 2002; Gesbert, Haumonte, Bolcskei, Krishnamoorthy & Paulraj, 2002). MIMO is an antenna system processing at both the transmitter and receiver to provide better performance and capacity without requiring extra bandwidth and power.

Another trend is the next-generation network that will be a multi-service, multi-access network of networks, providing a number of advanced services anywhere, anytime. Networked virtual environments (NVEs) may be considered another advanced service in the merging of multimedia computing and communication technologies. A wide range of exciting NVE applications may be foreseen, ranging from virtual shopping and entertainment (especially games and virtual communities) to medicine, collaborative design, architecture and education/training. One of the most popular groups of NVEs is collaborative virtual environments (CVEs), which enable multiple users' collaboration (Joslin, Di Giacomo & Magnenat-Thalman, 2004). Distributed Interactive Virtual Environments (DIVE) is one of the most prominent and mature NVEs developed within the academic world. It supports various multi-user CVE applications over the Internet. Key issues to be resolved include localization, scalability and persistence (Frecon, 2004).

Another important field of research is the use of the medium-voltage network for communication purposes, such as Internet access over the wall socket, voice-over IP (VoIP) and home entertainment (i.e., streaming audio and video at data rates in excess of 10 Mbps) (Gotz, 2004). The power line

communications offer a permanent online connection that is not expensive, since it is based on an existing electrical infrastructure. The development of appropriate power line communication (PLC) systems turns out to be an interesting challenge for the communications engineer.

A major roadblock to the widespread adoption of VoIP applications is that 911 operators are unable to view the numbers of callers using IP phones. VoIP service providers have had a hard time replicating this service, limiting the technology's usefulness in emergencies. Enhancements to VoIP services are being developed.

Next- or fourth-generation (NG/4G) wireless systems, currently in the design phase, are expected to support considerably higher data rates and will be based on IP technology, making them an integral part of the Internet infrastructure. Fourth-generation paradigm is combining heterogeneous networks, such as cellular wireless hot spots and sensor networks, together with Internet protocols. This heterogeneity imposes a significant challenge on the design of the network protocol stack. Different solutions include an adaptive protocol suite for next-generation wireless data networks (Akyildiz, Altunbasak, Fekri & Sivakumar, 2004) or evolution to cognitive networks (Mahonen, Rihujarvi, Petrova & Shelby, 2004), in which wireless terminals can automatically adapt to the environment, requirements and network.

One of the main goals for the future of telecommunication systems is service ubiquity (i.e., the ability for the user to transparently access any service, anywhere, anytime) based on a software reconfigurable terminal, which is part of ongoing European research activities in the context of reconfigurable software systems (Georganopoulos, Farnham, Burgess, Scholler, Sessler, Warr, Golubicic, Platbrood, Souville & Buljore, 2004). The use of mobile intelligent agents and policies is quite promising.

STANDARDS

All devices on a home network require a protocol and software to control the transmission of signals across the home network and Internet. A variety of standard protocols are installed in devices depending on the type of device. The TCP/IP suite of protocols is the standard protocol for linking computers on the Internet and is the fundamental building technology in home networks for entertainment services and Web applications. Currently, several companies and standardization groups are working on defining new protocols for the emerging technologies and interconnections with already defined protocols.

For example, the International Telecommunication Union and Institute of Electrical and Electronics Engineers (IEEE) is developing standards for passive optical networks (PON) capable of transporting Ethernet frames at gigabit-per-second speeds. The Ethernet Gigabit PON (GPON) system aligned with Full Services Access Network (FSAN)/ITU-T specification focuses on the efficient support of any level of Quality of Service (QoS). The Ethernet in the first mile (EFM) initiative of the IEEE and the GPON of FSAN/ITU-T solution represents a cost-effective solution for the last mile (Angelopoulos, Leligou, Argyriou, Zontos, Ringoot & Van Caenegem, 2004). Collaborative virtual environments (CVE) are being standardized by the Moving Picture Experts Group (MPEG). MPEG is one of the most popular standards for video and audio media today, while only a few years after its initial standardization. Recently, multiuser technology (MUTech) has been introduced to MPEG-4, Part 11, in order to provide some kind of collaborative experience using the MPEG specification. Although mobile voice services dominate the market, there is a need for more cellular bandwidth and new standards through General Packet Radio Service (GPRS) to third-generation wireless (3G) systems (Vriendt De, Laine, Lerouge & Xu, 2002). GPRS provides packet-switched services over the GSM radio and new services to subscribers. Creating ubiquitous computing requires seamlessly combining these wireless technologies (Chen, 2003). The Universal Mobile Telecommunication System (UMTS) is the chosen evolution of all GSM networks and Japanese Personal Digital Cellular network supporting IP-based multimedia.

More security specifications for wireless technologies (Farrow, 2003) are being developed. For example, Wired Equivalent Privacy (WEP) is improved with Wireless Protected Access (WPA). However, WPA is an interim standard that will be replaced with IEEE 802.11i standard upon its completion.

In addition to current developments, recent standards were specified or enhanced to be commercialized. IEEE ultrawideband (UWB) task group speci-

fied the UWB standard, which promises to revolutionize home media networking, with data rates between 100 and 500 Mbps. UWB could be embedded in almost every device that uses a microprocessor. For example, readings from electronic medical thermometers could automatically be input into the electronic chart that records vital statistics of a patient being examined. The UWB standard incorporates a variety of NG security mechanisms developed for IEEE 802.11 as well as plug-and-play features (Stroh, 2003). Another standard, IEEE 802.16 for wireless Metropolitan Access Networks (MANs), is commercialized by WiMax Forum, an industry consortium created to commercialize it, which allows users to make the connection between homes and the Internet backbone and to bypass their telephone companies (Testa, 2003). The IEEE 802.16a standard is a solution based on orthogonal frequency-division multiplexing, allowing for obstacle penetration and deployment of non line-of-sight (NLOS) scenarios (Koffman & Roman, 2002). Another example of enhancement is the DOCSIS 2.0 (Data Over Cable Service Interface Specifications) standard to provide the QoS capabilities needed for IP-specific types of broadband access, telephony and other multimedia applications provided by the cable industry.

CONCLUSION

Home networking presents novel challenges to systems designers (Teger, 2002). These include requirements such as various connection speeds, broadband access for the last mile (DSL, cable, fiber or wireless), current and future services, security, new applications oriented on home appliances, multiple home networks carrying multiple media (data, voice, audio, graphics, and video) interconnected by a backbone intranet, specific bandwidth requirements for different streams and so forth. Information Technology is moving toward digital electronics, and the major players in the industry will position for the future based on a functional specialization such as digitized content, multimedia devices and convergent networks. The information industry will realign to three main industries: Information Content, Information Appliances and Information Highways. These major paradigm shifts are coupled with changes from narrowband transmission to broadband communications and interactive broadband. The interactive broadband will have sociological implications on how people shop, socialize, entertain, conduct business and handle finances or health problems.

REFERENCES

Akyildiz, I., Altunbasak, Y., Fekri, F., & Sivakumar, R. (2004). AdaptNet: An adaptive protocol suite for the next-generation wireless Internet. *IEEE Communications Magazine, 42*(3), 128-136.

Angelopoulos, J.D., Leligou, H. C., Argyriou, T., Zontos, S., Ringoot, E., & Van Caenegem, T. (2004). Efficient transport of packets with QoS in an FSAN-aligned GPON. *IEEE Communications Magazine, 42*(2), 92-98.

Chen, Y-F.R. (2003). Ubiquitous mobile computing. *IEEE Internet Computing, 7*(2), 16-17.

Cherry, M. (2003). The wireless last mile. *IEEE Spectrum, 40*(9), 9-27.

DSL Forum Report. (2003). Retrieved from *www.dslforum.org/PressRoom/news_3.2.2004_EastEU.doc*

Farrow, R. (2003). Wireless security: Send in the clowns? *Network Magazine, 18*(9), 54-55.

Frecon, E. (2004). DIVE: Communication architecture and programming model. *IEEE Communications Magazine, 42*(4), 34-40.

Georganopoulos, N., Farnham, T., Burgess, R., Scholler, T., Sessler, J., Warr, P., Golubicic, Z., Platbrood, F., Souville, B., & Buljore, S. (2004). Terminal-centric view of software reconfigurable system architecture and enabling components and technologies. *IEEE Communications Magazine, 42*(5), 100-110.

Gesbert, D., Haumonte, L., Bolcskei, H., Krishnamoorthy, R., & Paulraj, A.J. (2002). Technologies and performance for non-line-of-sight broadband wireless access networks. *IEEE Communications Magazine, 40*(4), 86-95.

Gotz, M., Rapp, M., & Dostert, K. (2004). Power line channel characteristics and their effect on communication system design. *IEEE Communications Magazine, 42*(4), 78-86.

Hentea, M. (2004). Data mining descriptive model for intrusion detection systems. *Proceedings of the 2004 Information Resources Management Association International Conference,* (pp. 1118-1119). Hershey, PA: Idea Group Publishing.

Joslin, C., Di Giacomo, T., & Magnenat-Thalman, N. (2004). Collaborative virtual environments: From birth to standardization. *IEEE Communications Magazine, 42*(4), 28-33.

Koffman, I., & Roman, V. (2002). Broadband wireless access solutions based on OFDM access in IEEE 802.16. *IEEE Communications Magazine, 40*(4), 96-103.

Mahonen, P., Rihujarvi, J., Petrova, M., & Shelby, Z. (2004). Hop-by-hop toward future mobile broadband IP. *IEEE Communications Magazine, 42*(3), 138-146.

Murch, R.D., & Letaief, K.B. (2002). Antenna systems for broadband wireless access. *IEEE Communications Magazine, 40*(4), 76-83.

Stamper, D.A., & Case, T.L. (2003). *Business data communications* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Stroh, S. (2003). Ultra-wideband: Multimedia unplugged. *IEEE Spectrum, 40*(9), 23-27.

Teger, S., & Waks, D.J. (2002). End-user perspectives on home networking. *IEEE Communications Magazine, 40*(4), 114-119.

Testa, B.M. (2003). U.S. phone companies set stage for fiber to the curb. *IEEE Spectrum, 40*(9), 14-15.

Vriendt De, J., Laine, P., Lerouge, C., & Xu, X. (2002). Mobile network evolution: A revolution on the move. *IEEE Communications Magazine, 40*(4), 104-111.

Zhang, M., & Wolff, R.S. (2004). Crossing the digital divide: Cost effective broadband wireless access for rural and remote areas. *IEEE Communications Magazine, 42*(2), 99-105.

B

KEY TERMS

Broadband Access: A form of Internet access that provides information and communication services to end users with high-bandwidth capabilities.

Broadband Transmission: A form of data transmission in which data are carried on high-frequency carrier waves; the carrying capacity medium is divided into a number of subchannels for data such as video, low-speed data, high-speed data and voice, allowing the medium to satisfy several communication needs.

Broadband Wireless Access: A form of access using wireless technologies.

Cable Access: A form of broadband access using a cable modem attached to a cable TV line to transfer data.

Digital Subscriber Line (DSL): A technique for transferring data over regular phone lines by using a frequency different from traditional voice calls or analog modem traffic over the phone wires. DSL lines carry voice, video and data.

MPEG-4: Standard specified by Moving Picture Experts Group (MPEG) to transmit video and images over a narrower bandwidth and can mix video with text, graphics and 2-D and 3-D animation layers.

Challenges and Perspectives for Web-Based Applications in Organizations

George K. Lalopoulos
Hellenic Telecommunications Organization S.A. (OTE), Greece

Ioannis P. Chochliouros
Hellenic Telecommunications Organization S.A. (OTE), Greece

Anastasia S. Spiliopoulou-Chochliourou
Hellenic Telecommunications Organization S.A. (OTE), Greece

INTRODUCTION

The last decade is characterized by the tempestuous evolution, growth, and dissemination of information and communication technologies in the economy and society. As a result, information and communication technologies have managed to open new broad horizons and markets for enterprises, whose installation and operation costs are rapidly amortizing through the proper usage of these technologies.

The most common systems used are cellular phones, stand-alone PCs (Personal Computers), networks of PCs, e-mail and EDI (Electronic Data Interchange), Personal Digital Assistants (PDAs), connection to the Internet, and corporate Web pages. In particular, the evolution in speed, operability, and access to the World Wide Web, and the penetration of e-commerce in conjunction with the internationalization of competition have set up new challenges as well as perspectives for enterprises: from small and medium sized to large ones. Even very small enterprises—with up to nine employees—have conceived the importance of Internet access, and a considerable percentage of them have access to the Web.

In today's worldwide environment, markets tend to become electronic, and national boundaries, as far as markets are concerned, tend to vanish. Buyers can find a product or service through the Internet at a lower price than that of a local market. Enterprises, on the other hand, can use the Internet in order to extend their customer basis and at the same time organize more efficiently the communication with their suppliers and partners. Thus, enterprises can reduce costs, increase productivity, and surpass the narrow geographical boundaries, enabling cooperation with partners from different places and countries. One memorable example is the company Amazon.com, which introduced the offering of books through its Web page, thus leaving behind the traditional bookstore. In addition, enterprises can use the new information and communication technologies so as to organize and coordinate the internal communication of their various departments as well as their structure more efficiently, taking into account factors like business mobility and distribution.

These demands have caused many companies to consider the convergence of voice, video, and data through IP- (Internet Protocol) centric solutions that include rich and streaming media together with various IP-based applications such as VoIP (Voice Over Internet Protocol), video, messaging, and collaboration as a way to lower costs and deliver product-enhancing applications to end users in a secure environment. However, it is not always easy for a company to keep pace with innovation due to financial restrictions and internal politics.

TODAY'S IT CHALLENGES

Today's CIOs (Chief Information Officers) face higher expectations. Some of the most significant challenges are the following (Pandora Networks, 2004).

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Using Technology to Increase Productivity

New applications offer a standard, open platform providing for new communications features, such as voice services (Intelligent Call Routing [ICR], Unified Messaging [UM], etc.) and non voice services (Instant Messaging [IM], Web collaboration, conferencing, etc.). These features make users more productive by streamlining their communications and access to information. Moreover, Web-based administration provides for simpler management and quicker response of the technical staff to end users. The latter also have the possibility to manage their services.

Servicing an Increasingly Mobile and Distributed Workforce

As the workforce become less centralized and static, unified communications enable IT to deliver the same functionality to the remote office as the corporate headquarters. Mobile and distant users can access the same applications as their colleagues at the headquarters. They can also communicate with other users as if they were in the same location.

Delivering Revenue-Generating Applications and Features

Unified communications provide a foundation for future revenue-generating applications. For example, a new customer-support application will provide for a higher level of real-time customer interaction by enabling customers to have access to trained service engineers that can resolve problems with IP-based interaction tools. This improves customer service and enhances customer loyalty and long-term value. As another example, multimedia applications can enable collaboration, shortening project life cycles.

Reducing Costs

By managing one converged infrastructure, IT departments can reduce administrative and management complexity, therefore reducing all related costs. If an employee is moving, the same person that relocated the PC can also move the phone. Using a client-server architecture, end-user telephones become plug and play. Convergence also offers the opportunity to introduce new applications that can be used to replace expensive metered services (e.g., IP conferencing could be used instead of conference calls).

Unifying All Communications Platforms

With unified communications, users can access corporate information from any device, regardless of the underlying platform. A common user often has five or six different communication services (e.g., phones, fax, e-mail, etc.), each performing the same basic function, that is, contacting the user. By reducing the number of contact methods from five or six to just one, unified communications reduces complexity, increases productivity and responsiveness, and enables collaboration.

Aligning IT and Business Processes

Convergence delivers an open and integrated communications platform that gives CIOs the opportunity to optimize existing business processes. For example, corporate directories could be integrated into IP phones and other collaboration tools, enabling end users to access all corporate information from multiple devices. As a result, the ability to reinvest business processes, drive down costs, and deliver value to the company is enhanced. Furthermore, optimization software tools and decision-support systems can be combined with Web-service technology to deliver distributed applications to users as remote services. Optimization models are considered as critical components for an organization's analytical IT systems as they are used to analyze and control critical business measures such as cost, profit, quality, and time. One can think of modeling and solver tools as the components offered by the provider with added infrastructure consisting of secure data storage and data communication, technical support on the usage of tools, management and consultancy for the development of user-specific models and applications, and some measure of the quality of the provided optimization services. Applications include sectors like finance, manufacturing and supply-chain management, energy and utilities, and environmental planning. The OPT (Optimization Service Provider; http://www.ospcraft.com/) and WEBOPT (Web-Enabled Optimiza-

tion Tools and Models for Research, Professional Training, and Industrial Deployment; http://www.webopt.org/) projects are based on this view (Valente & Mitra, 2003).

EXISTING TECHNOLOGIES

Considering that the Web is fundamentally a new medium of human communication, not just a technology for information processing or computation, its evolution depends on media design, and Web services and applications design. Media design has evolved into rich media, which is a method of communication used for performing enterprise core business operations, such as real-time corporate communication, e-learning, sales training, marketing (e.g., online advertising), and collaboration, that comprises animation, sound, video, and/or interactivity. It is deployed via standard Web and wireless applications (Little, 2004). Rich media typically allows users to view and interact with products or services. It includes standard-sized banners with forms or pull-down, pop-up, or interstitial menus; streaming audio and video; animation plug-ins; and so forth. Text, as well as standard graphic formats such as JPEG (Joint Photographic Experts Group) and GIF (Graphics Interchange Format), would not be considered rich media. Broadband technology enables both content providers and enterprises to create more rich-media-based content (Adverblog.com, 2004).

Advanced Web services and applications offer an attractive platform for business applications and organizational information systems. They offer capabilities such as chat, Web collaboration, presentation sharing, streaming video delivery to various locations, and so forth, thus enhancing cooperation and productivity in a distributed environment. Furthermore, Web technology is often presented as a revolution in network and information technologies, propelling change from static, hierarchical structures to more dynamic, flexible, and knowledge-based organizational forms. Current research efforts are oriented toward interactive Web applications that mediate interaction among multiple distributed actors who are not only users but also designers in the sense that they contribute to the system's structure and content (Valente & Mitra, 2003).

SECURITY

The growing use of the Internet by organizations for transactions involving employees, business partners, suppliers, and customers has led to a need for increased security demands in order to protect and preserve private resources and information. Moreover, Web security becomes more significant as the amount of traffic through the Internet is increasing and more important transactions are taking place, especially in the e-commerce domain (Shoniregun, Chochliouros, Lapeche, Logvynovskiy, & Spiliopoulou-Chochliourou, 2004).

The Internet has become more dangerous over the last few years with specific network-security threats such as the following.

- Faults in servers (OS [Operating System] bugs, installation mistakes): the most common holes utilized by hackers
- Weak authentication
- Hijacking of connections (especially with unsecured protocols)
- Interference threats such as jamming and crashing the servers using, for example, Denial of Service (DoS) attacks
- Viruses with a wide range of effects
- Active content with Trojans
- Internal threats

In order to deal with this matter, preventive measures, such as the use of firewalls (implemented in hardware, software, or both) or data encryption for higher levels of security, are taken. One interesting approach to support Web-related security in an organization, especially in an extranet-like environments, is the use of Virtual Private Networks (VPNs) based on a choice of protocols, such as IPsec (IP Security Protocol) and Secure-Sockets Layer (SSL).

IPsec refers to a suite of Internet Engineering Task Force (IETF) protocols that protect Internet communications at the network layer through encryption, authentication, confidentiality, antireplay protection, and protection against traffic-flow analysis at the network layer. IPsec VPNs require special-purpose client software on the remote user's access device to control the user side of the communication link (Nortel Networks, 2002). This requirement

makes it more difficult to extend secure access to mobile users, but it increases VPN security by ensuring that access is not opened from insecure computers (such as PCs at public kiosks and Internet cafes). IPsec implementation is a time-consuming task, usually requiring changes to the firewall, the resolution of any NAT (Network Address Translation) issues, and the definition of sophisticated security policies to ensure users have access only to permitted areas on the network. Thus, IPsec VPNs are a good fit for static connections that do not change frequently.

The SSL protocol uses a private key (usually 128 bits) to encrypt communications between Web servers and Web browsers for tunneling over the Internet at the application layer. Therefore, a certificate is needed for the Web server. SSL support is built into standard Web browsers (Internet Explorer, Netscape Navigator, etc.) and embedded into a broad range of access devices and operating systems. SSL is suitable for remote users needing casual or on-demand access to applications such as e-mail and file sharing from diverse locations (such as public PCs in Internet kiosks or airport lounges) provided that strong authentication or access-control mechanisms are enacted to overcome the inherent risks of using insecure access devices (Viega & Messier, 2002).

ROI (Return on Investment) is one of the most critical areas to look at when analyzing SSL vs. IPsec VPN. Lower telecommunication costs, reduced initial implementation costs, substantially decreased operational and support costs, easy user scaling, open user access, and ease of use have rendered SSL the most widely used protocol for securing Web connections (Laubhan, 2003). However, there are some problems regarding SSL. The key-generation process requires heavy mathematics depending on the number of bits used, therefore increasing the response time of the Web server. After the generation of the key pair, an SSL connection is established. As a consequence, the number of connections per second is limited and fewer visitors can be served when security is enabled. Moreover, with increased delays in server response, impatient users will click the reload button on their browsers, initiating even more SSL connection requests when the server is most busy. These problems can be handled with techniques like choosing the right hardware and software architecture (e.g., SSL accelerator), designing graphics and composite elements as a single file rather than "sliced" images, and so forth (Rescorla, 2000).

Possible security loopholes come from the fact that SSL is based on the existence of a browser on the user side. Thus, the browser's flaws could undermine SSL security. Internet Explorer, for example, has a long history of security flaws, the vast majority of which have been patched. The heterogeneity of Web clients to offer service to a wide range of users and devices also creates possible security loopholes (e.g., the risk of automatic fallback to an easily cracked 40-bit key if a user logs in with an outdated browser). Other loopholes result from the fact that many implementations use certificates associated with machines rather than users. The user who leaves machines logged-in and physically accessible to others, or the user who enables automatic-login features makes the security of the network depend on the physical security that protects the user's office or, worse still, the user's portable device. According to an academic report from Darmouth College (Ye, Yuan, & Smith, 2002), no solution is strong enough to be a general solution for preventing Web spoofing. However, ongoing research is expected to decrease browser vulnerability.

COMMERCIAL PRODUCTS

Some indicative commercial products are the following:

- Spanlink (http://www.spanlink.com/) offers the Concentric Solutions Suite that comprises a number of products (Concentric Agent, Concentric Supervisor, Concentric Customer, etc.).

The aim of these products is to optimize the way customers interact with businesses over the Internet and over the phone.

For example, with Concentric Agent, agents can readily access an interface with features such as an automated screen pop of CRM (Customer Relationship Management) and help-desk applications, a highly functioning soft-phone toolbar, real-time statistics, and chat capabilities.

Concentric Customer provides automated self-service options for customers over the phone and over the Web, making it easy to be successful in finding precise answers on a company's Web site.

Concentric Supervisor (EETIMES.com, 2003) focuses on supervisors and their interaction with agents. It integrates real-time visual and auditory monitoring, agent-to-supervisor chat capabilities, and call-control functions.

- Convoq (http://www. convoq.com) offers Convoq ASAP, a real-time, rich-media instant-messaging application: the intimacy of videoconferencing and the power of Web conferencing to meet all the collaboration needs in a company. Through its use, participants can obtain services including chat, broadcast audio and video, and the sharing of presentations with the use of Windows, Macintosh, or Linux Systems without the need for downloads or registrations. ASAP supports SSL (Convoq Inc., 2004).
- Digital Media Delivery Solution (DMDS; Kontzer, 2003) is a digital media solution offered by the combined forces of IBM, Cisco Systems, and Media Publisher Inc. (MPI; http://www.media-publisher. com/). It allows any organization in any industry to quickly and efficiently deliver rich media including streaming video to geographically dispersed locations. It is designed to provide streaming technology that helps customers leverage digital media in every phase of their business life cycle.
- BT Rich Media (British Telecom, 2004) is a new digital media platform designed to provide tools to allow businesses (especially content providers) and individuals to create and distribute digital content on the Web. It was developed by BT in partnership with Real Networks and TWI (Trans World International). The launch of the product on April 6, 2004, was in line with BT's strategy to reach its target of broadband profitability by the end of 2005, as well as fighting off increasing pressure from broadband competitors.

FUTURE TRENDS

The current Web is mainly a collection of information but does not yet provide adequate support in processing this information, that is, in using the computer as a computational device. However, in a business environment, the vision of flexible and autonomous Web service translates into automatic cooperation between enterprise services. Examples include automated procurement and supply-chain management, knowledge management, e-work, mechanized service recognition, configuration, and combination (i.e., realizing complex work flows and business logics with Web services, etc.). For this reason, current research efforts are oriented toward a semantic Web (a knowledge-based Web) that provides a qualitatively new level of service. Recent efforts around UDDI (Universal Description, Discovery, and Integration), WSDL (Web-Services Description Language), and SOAP (Simple Object-Access Protocol) try to lift the Web to this level (Valente & Mitra, 2003).

Automated services are expected to further improve in their capacity to assist humans in achieving their goals by understanding more of the content on the Web, and thus providing more accurate filtering, categorization, and searches of information sources. Interactive Web applications that mediate interaction among multiple distributed designers (i.e., users contributing to the system's structure and content) is the vision of the future (Campell, 2003).

CONCLUSION

In today's competitive environment, enterprises need new communication applications that are accessible on devices located either at their premises or in remote locations. The Internet and the proliferation of mobile devices, such as mobile phones and wireless PDAs, are playing a very important role in changing the way businesses communicate. Free sitting (that is, the ability of an employee to sit in any office and use any PC and any phone to access his or her personalized working environment and to retrieve applications, messages, etc.), mobility, responsiveness, customer satisfaction, and cost optimization are key challenges that enterprises are facing today (Sens, 2002). Advanced technologies, such as rich and streaming media and new Web applications and services, are being used to develop a new generation of applications and services that can help businesses face these challenges.

One of the biggest challenges for businesses will be the ability to use teamwork among people in order

to network the entire knowledge of the company with the objective of providing first-class services to customers and developing innovative products and services. However, companies often face problems in adopting these new technologies, due mainly to bandwidth limitations, economic restrictions, and internal politics leading to hesitations and serious doubts about the return on investments.

Within the next three to five years, the increase by several orders of magnitude in backbone bandwidth and access speeds, stemming from the deployment of IP and ATM (Asynchronous Transfer Mode), cable modems, Radio Local Area Networks (RLANs), and Digital Subscriber Loop (DSL) technologies, in combination with the tiering of the public Internet in which users will be required to pay for the specific service levels they require, is expected to play a vital role in the establishment of an IP-centric environment. At the same time, Interactive Web applications among multiple distributed users and designers contributing to the system's structure and content, in combination with optimization tools and decision-support systems, are expected to change organizational structures to more dynamic, flexible, and knowledge-based forms.

REFERENCES

Adverblog.com. (2004). *Rich media archives: Broadband spurring the use of rich media.* Retrieved July 26, 2004, from *http:/www.aderblog.com/archives/cat_rich_media.htm*

British Telecom. (2004). *BT news: BT takes broadband revolution into new territory.* Retrieved July 28, 2004, from *http://www.btplc.com/News/Pressreleasesandarticles/Agencynewsreleases/2004/an0427.htm*

Campbell, E. (2003). Creating accessible online content using Microsoft Word. *Proceedings of the Fourth Annual Irish Educational Technology Conference.* Retrieved July 28, 2004, from *http://ilta.net/EdTech2003/papers/ecampbell_accessibility_word.pdf*

Convoq Inc. (2004). *ASAP security overview.* Retrieved July 27, 2004, from *http://www.convoq.com/Whitepapers/asapsecurity.pdf*

EETIMES.com. (2003). *Spanlinks's concentric supervisor product receivers Technology Marketing Corporation's TMC Labs innovation award 2002.* Retrieved July 27, 2004, from *http://www.eetimes.com/pressreleases/bizwire/42710*

Grigonis, R. (2004, March-April). Web: Everything for the enterprise. *Von Magazine, 2*(2). Retrieved July 27, 2004, from *http://www.vonmag.com/issue/2004/marapr/features/web_everything.htm*

Kontzer, T. (2003). IBM and Cisco team on digital media. *Information Week*. Retrieved July 28, 2004, from *http://www.informationweek.com/story/showArticle.jhtml?articleID=10100408*

Laubhan, J. (2003). *SSL-VPN: Improving ROI and security of remote access, rainbow technologies.* Retrieved July 28, 2004, from *http://www.mktg.rainbow.com/mk/get/SSL%20VPN%20-%20Improving%20ROI%20and%20Security%20of%20Remote%20Access.pdf*

Little, S. (2004). Rich media is… (Part 1 of 2). *Cash Flow Chronicles, 50*. Retrieved July 26, 2004, from *http://www.cashflowmarketing.com/newsletter*

Nortel Networks. (2002). *IPsec and SSL: Complementary solutions. A shorthand guide to selecting the right protocol for your remote-access and extranet virtual private network* (White paper). Retrieved July 27, 2004, from *http://www.nortelnetworks/com/solutions/ip_vpn/collateral/nn102260-110802.pdf*

Pandora Networks. (2004). *The business case of unified communications. Using worksmart IP applications, part one: Voice and telephony.* Retrieved July 26, 2004, from *http://www.pandoranetworks.com/whitepaper1.pdf*

Rescorla, E. (2000). *SSL and TLS: Designing & building secure systems.* Boston, Addison-Wesley.

Sens, T. (2002). Next generation of unified communications for enterprises: Technology White Paper. *Alcatel Telecommunications Review,* 4th quarter. Retrieved July 28, 2004, from *http://www.alcatel.com/doctypes/articlepaperlibrary/pdf/ATR2002Q4/T0212-Unified_Com-EN.pdf*

Shoniregun, C. A., Chochliouros, I. P., Lapeche, B., Logvynovskiy, O., & Spiliopoulou-Chochliourou, A.

S. (2004). *Questioning the boundary issues of Internet security.* London: e-Centre for Infonomics.

Valente, P., & Mitra, G. (2003). *The evolution of Web-based optimisation: From ASP to e-services* (Tech Rep. No. CTR 08/03). London: Brunel University, Department of Mathematical Sciences & Department of Economics and Finance, Centre for the Analysis of Risk and Optimisation Modelling Applications: CARISMA. Retrieved July 27, 2004, from *http://www.carisma.brunnel.ac.uk/papers/option_eservices_TR.pdf*

Viega, J., Messier, M., & Chandra, P. (2002). *Network security with open SSL.* Sebastopol, CA: O'Reilly & Associates.

Ye, E., Yuan, Y., & Smith, S. (2002). *Web spoofing revisited: SSL and beyond* (Tech. Rep. No. TR2002-417). Darmouth College, New Hampshire, USA, Department of Computer Science. Retrieved July 28, 2004, from *http://www.cs.darmouth.edu/pkilab/demos/spoofing*

KEY TERMS

Banner: A typically rectangular advertisement placed on a Web site either above, below, or on the sides of the main content and linked to the advertiser's own Web site. In the early days of the Internet, banners were advertisements with text and graphic images. Today, with technologies such as Flash, banners have gotten much more complex and can be advertisements with text, animated graphics, and sound.

ICR (Intelligent Call Routing): A communications service that provides companies with the ability to route inbound calls automatically to destinations such as a distributed network of employees, remote sites, or call-center agents. Call routing is typically based on criteria such as area code, zip code, caller ID, customer value, previous customer status, or other business rules.

IM (Instant Messaging): A type of communications service that enables you to create a kind of private chat room with another individual in order to communicate in real time over the Internet. It is analogous to a telephone conversation but uses text-based, not voice-based, communication. Typically, the instant-messaging system alerts you whenever somebody on your private list is online. You can then initiate a chat session with that particular individual.

Interstitial: A page that is inserted in the normal flow of the editorial content structure on a Web site for the purpose of advertising or promoting. It is usually designed to move automatically to the page the user requested after allowing enough time for the message to register or the advertisement(s) to be read.

SOAP (Simple Object-Access Protocol): A lightweight XML- (extensible markup language) based messaging protocol used to encode the information in Web-service request and response messages before sending them over a network. SOAP messages are independent of any operating system or protocol and may be transported using a variety of Internet protocols, including SMTP (Simple Mail Transfer Protocol), MIME (Multipurpose Internet Mail Extensions), and HTTP (Hypertext Transfer Protocol).

UDDI (Universal Description, Discovery, and Integration): A directory that enables businesses to list themselves on the Internet and discover each other. It is similar to a traditional phone book's yellow and white pages.

UM (Unified Messaging): It enables access to faxes, voice mail, and e-mail from a single mailbox that users can reach either by telephone or a computer equipped with speakers.

WSDL (Web-Services Description Language): An XML-formatted language used to describe a Web service's capabilities as collections of communication endpoints capable of exchanging messages. WSDL is an integral part of UDDI, an XML-based worldwide business registry. WSDL is the language that UDDI uses. WSDL was developed jointly by Microsoft and IBM.

Collaborative Web-Based Learning Community

Percy Kwok Lai-yin
Chinese University of Hong Hong, China

Christopher Tan Yew-Gee
University of South Australia, Australia

INTRODUCTION

Because of the ever-changing nature of work and society under the knowledge-based economy in the 21st century, students and teachers need to develop ways of dealing with complex issues and thorny problems that require new kinds of knowledge that they have never learned or taught (Drucker, 1999). Therefore, they need to work and collaborate with others. They also need to be able to learn new things from a variety of resources and people and investigate questions, then bring their learning back to their dynamic life communities. There have arisen in recent years *learning-community* approaches (Bereiter, 2002; Bielaczyc & Collins, 1999) and *learning-ecology* (Siemens, 2003) or *information-ecology* approaches (Capurro, 2003) to education. These approaches fit well with the growing emphasis on lifelong, life-wide learning and knowledge-building works.

Following this trend, Internet technologies have been translated into a number of strategies for teaching and learning (Jonassen, Howland, Moore, & Marra, 2003) with supportive development of one-to-one (e.g., e-mail posts), one-to-many (such as e-publications), and many-to-many communications (like videoconferencing). The technologies of computer-mediated communications (CMC) make online instruction possible and have the potential to bring enormous changes to student learning experiences in the real world (Rose & Winterfeldt, 1998). It is because individual members of learning communities or ecologies help synthesize learning products via deep information processing, mutual negotiation of working strategies, and deep engagement in critical thinking, accompanied by an ownership of team works in those communities or ecologies (Dillenbourg, 1999). In short, technology in communities is essentially a means of creating fluidity between knowledge segments and connecting people in learning communities. However, this Web-based collaborative learning culture is neither currently emphasized in local schools nor explicitly stated out in intended school-curriculum guidelines of formal educational systems in most societies. More than this, community ownership or knowledge construction in learning communities or ecologies may still be infeasible unless values in learning cultures are necessarily transformed after the technical establishment of Web-based learning communities.

BACKGROUND

Emergence of a New Learning Paradigm through CMC

Through a big advance in computer-mediated technology (CMT), there have been several paradigm shifts in Web-based learning tools (Adelsberger, Collis, & Pawlowski, 2002). The first shift moves from a *content-oriented* model (information containers) to a *communication-based* model (communication facilitators), and the second shift then elevates from a communication-based model to a *knowledge-construction* model (creation support). In the knowledge-construction model, students in a Web-based discussion forum mutually criticize each other, hypothesize pretheoretical constructs through empirical data confirmation or falsification, and with scaffolding supports, coconstruct new knowledge beyond their existing epistemological boundaries under the social-constructivism paradigm (Hung, 2001). Noteworthy is the fact that the knowledge-construction model can only nourish a learning community or ecology, and it is advocated by some

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

cognitive scientists in education like Collins and Bielaczyc (1997) and Scardamalia and Bereiter (2002). Similarly, a Web-based learning ecology contains intrinsic features of a collection of overlapping communities of mutual interests, cross-pollinating with each other and constantly evolving with largely self-organizing members (Brown, Collins, & Duguid, 1989), in the knowledge-construction model.

Scaffolding Supports and Web-Based Applications

According to Vygotsky (1978), the history of the society in which a child is reared and the child's personal history are crucial determinants of the way in which that individual will think. In this process of cognitive development, language is a crucial tool for determining how the child will learn how to think because advanced modes of thought are transmitted to the child by means of words (Schütz, 2002). One essential tenet in Vygotsky's theory is the notion of the existence of what he calls the zone of proximal development (ZPD). The child in this *scaffolding* process of ZPD, providing nonintrusive intervention, can be an adult (parent, teacher, caretaker, language instructor) or another peer who has already mastered that particular function. Practically, the scaffolding teaching strategy provides individualized supports based on the learner's ZPD. Notably, the scaffolds facilitate a student's ability to build on prior knowledge and internalize new information. The activities provided in scaffolding instruction are just beyond the level of what the learner can do alone. The more capable peer will provide the scaffolds so that the learner can accomplish (with assistance) the tasks that he or she could otherwise not complete, thus fostering learning through the ZPD (Van Der Stuyf, 2002).

In Web-based situated and anchored learning contexts, students have to develop metacognition to learn how, what, when, and why to learn in genuine living contexts, besides problem-based learning contents and methods in realistic peer and group collaboration contexts of synchronous and asynchronous interactions. Empirical research databases illuminate that there are several levels of Web uses or knowledge-building discourses ranging from mere informational stages to coconstruction stages (Gilbert, & Driscoll, 2002; Harmon & Jones, 2001). To sum up, five disintegrating stages of Web-based learning communities or ecologies are necessarily involved in Table 1.

Noteworthy is that the students succeed in developing scaffold supports via ZPD only when they attain coconstruction levels of knowledge construction, at which student-centered generation of discussion themes, cognitive conflicts with others' continuous critique, and ongoing commitments to the learning communities (by having constant attention and mutual contributions to discussion databases) are emerged. It should be noted that Web-based discussion or sharing in e-newsgroups over the Internet may not lead to communal ownership or knowledge construction.

Key Concepts of Communities of Practice

Unlike traditional static, lower order intelligence models of human activities in the Industrial Age, new higher order intelligence models for communities of practice have emerged. Such models are complex-adaptive systems, employing self-organized, free-initiative, and free-choice operating principles, and creating human ecology settings and stages for its acting out during the new Information Era. Under the technological facilitation of the Internet, this new emerging model is multicentered, complex adaptive, and self-organized, founded on the dynamic human relationships of equality, mutual respect, and deliber-

Table 1. Five disintegrating stages of Web-based learning communities

Disintegrating stages	Distinctive Features
Informational Level	Mere dissemination of general information
Personalized Level	Members' individual ownership in the communities
Communicative Level	Members' interactions found in the communities
Communal Level	Senses of belonging or communal ownership built up
Co-construction Level	Knowledge-construction among members emerged

ate volition. When such a model is applied to educational contexts, locally managed, decentralized marketplaces of lifelong and life-wide learning take place. In particular, teacher-student partnerships are created to pursue freely chosen and mutually agreed-upon learning projects (Moursund, 1999), and interstudent coconstruction of knowledge beyond individual epistemological boundaries are also involved (Lindberg, 2001). Working and learning are alienated from one another in formal working groups and project teams; however, communities of practice and informal networks (embracing the above term Web-based learning communities) both combine working and knowledge construction provided that their members have a commitment to the professional development of the communities and mutual contributions to generate knowledge during collaborations. In particular, their organization structures can retain sustainability even if they lose active members or coercive powers (Wenger, McDermott, & Snyder, 2002). It follows that students engaging in communities of practice can construct knowledge collaboratively when doing group work.

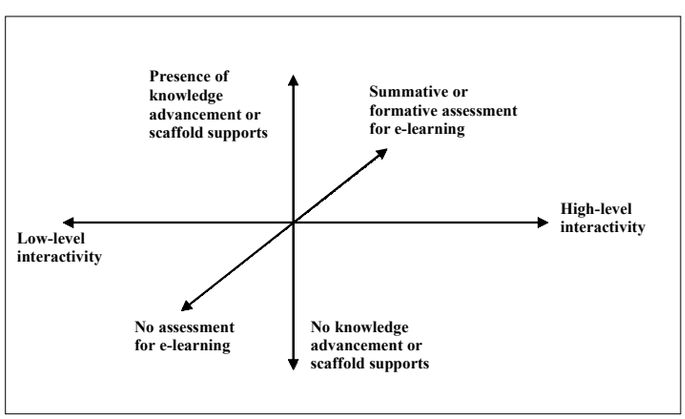
Main Focus of the Paper

In learning-community or -ecology models, there arise some potential membership and sustainability problems. Despite their technical establishments, some Web-based learning ecologies may fail to attain the communal or coconstruction stages (see Table 1), or fail to sustain after their formation.

Chan, Hue, Chou, and Tzeng (2001) depict four spaces of learning models, namely, the future classroom, the community-based, the structural-knowledge, and the complex-problem learning models, which are designed to integrate the Internet into education. Furthermore, Brown (1999, p. 19) points out that "the most promising use of Internet is where the buoyant partnership of people and technology creates powerful new online learning communities." However, the concept of communal membership is an elusive one. According to Slevin (2000, p. 92), "It might be used to refer to the communal life of a sixteenth-century village—or to a team of individuals within a modern organization who rarely meet face to face but who are successfully engaged in online collaborative work."

To realize cognitive models of learning communities, *social communication* is required since human effort is the crucial element. However, the development of a coercive learning community at the communal or coconstruction levels (Table 1) is different from the development of a social community at the communicative level, though "social communication is an essential component of educational activity" (Harasim, 1995). Learning communities are complex systems and networks that allow adaptation and change in learning and teaching practices (Jonassen, Peck, & Wilson, 1999). Collins and Bielaczyc (1997) also realize that knowledge-building communities require sophisticated elements of *cultural transformation* while Gilbert and Driscoll (2002) observe that learning quantity and quality

Figure 1. A three-dimensional framework for classifying Web-based learning


Presence of knowledge advancement or scaffold supports
Summative or formative assessment for e-learning
High-level interactivity
Low-level interactivity
No assessment for e-learning
No knowledge advancement or scaffold supports

depend on the value beliefs, expectations, and learning attitudes of the community members. It follows that some necessary conditions for altering basic educational assumptions held by community learners and transforming the entire learning culture need to be found out for epistemological advancement. On evaluation, there are three intrinsic dimensions that can advance students' learning experiences in Web-based learning communities. They are the degree of interactivity, potentials for knowledge construction, and assessment of e-learning.

For the systematic classification of Web-based learning communities, a three-dimensional conceptual framework is necessarily used to highlight the degree of interactivity (one-to-one, one-to-many, and many-to-many communication modes), presence or absence of scaffolding or knowledge-advancement tools (coconstruction level in Table 1), and modes of learning assessments (no assessment, summative assessment for evaluating learning outcomes, or formative assessment for evaluating learning processes; Figure 1).

This paper provides some substantial knowledge-construction aspects of most collaborative Web-based learning communities or learning ecologies. Meantime, it conceptualizes the crucial sense of scaffolding supports and addresses underresearched sociocultural problems for communal membership and knowledge construction, especially in Asian school curricula.

FUTURE TRENDS

The three issues of cultural differences, curricular integration, and leadership transformation in Web-based learning communities are addressed here for forecasting their future directions. Such collaborative Web-based learning communities have encountered the sociocultural difficulties of not reaching group consensus necessarily when synthesizing group notes for drawing conclusions (Scardamalia, Bereiter, & Lamon, 1995). Other sociocultural discrepancies include the following (Collins & Bielaczyc, 1997; Krechevsky & Stork, 2000; Scardamalia & Bereiter, 1996).

- discontinuous expert responses to students' questions, thereby losing students' interest
- students' overreliance on expert advice instead of their own constructions
- value disparities in the nature of collaborative discourses between student construction and expertise construction of knowledge

The first issue is influenced by the heritage culture upon Web-based learning communities or ecologies. Educational psychologists (e.g., Watkins & Biggs, 2001) and sociologists (e.g., Lee, 1996) also speculate the considerable influence of the heritage of Chinese culture upon the roles of teachers and students in Asian learning cultures. When knowledge building is considered as a way of learning in Asian societies under the influence of the heritage of Chinese culture, attention ought to be paid to teachers' as well as students' conceptions, and Asian cultures of learning and teaching, especially in a CMC learning community.

The second issue is about curricular integration. There come some possible cases in which participating teachers and students are not so convinced by CMC or do not have a full conception of knowledge building when establishing collaborative learning communities. More integration problems may evolve when school curricula are conformed to the three pillars of conventional pedagogy, namely, reduction to subject matter, reduction to activities, and reduction to self-expression (Bereiter, 2002). Such problems become more acute in Asian learning cultures, in which there are heavy stresses on individually competitive learning activities, examination-oriented school assessments, and teacher-led didactical interactions (Cheng, 1997).

The third issue is about student and teacher leadership in cultivating collaborative learning cultures (Bottery, 2003). Some preliminary sociocultural research findings (e.g., Yuen, 2003) reveal that a high sense of membership and the necessary presence of proactive teacher and student leaders in inter- and intraschool domains are crucial for knowledge building in Web-based learning communities or ecologies.

CONCLUSION

To sum up, there are some drawbacks and sociocultural concerns toward communal membership, knowl-

edge-construction establishment, and the continuation of learning ecologies (Siemens, 2003).

- lack of internal structures for incorporating flexibility elements
- inefficient provision of focused and developmental feedback during collaborative discussion
- no directions for effective curricular integration for teachers' facilitation roles
- no basic mechanisms of pinpointing and eradicating misinformation or correcting errors in project works
- lack of assessment for evaluating learning processes and outcomes of collaborative learning discourses

So there comes an urgent need to address new research agendas to investigate the shifting roles of students and teachers (e.g., at the primary and secondary levels) and their reflections on knowledge building, and to articulate possible integration models for project works in Asian school curricula with high student-teacher ratios and prevalent teacher-centered pedagogy when Web-based learning communities or ecologies are technically formed.

REFERENCES

Adelsberger, H. H., Collis, B., & Pawlowski, J. M. (Eds.). (2002). *Handbook on information technologies for education and training.* Berlin & Heidelberg, Germany: Springer-Verlag.

Bereiter, C. (2002). *Education and mind in the knowledge age.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bielaczyc, K., & Collins, A. (1999, February). Learning communities in classroom: Advancing knowledge for a lifetime. *NASSP Bulletin*, 4-10.

Bottery, M. (2003). The leadership of learning communities in a culture of unhappiness. *School Leadership & Management, 23*(2), 187-207.

Brown, J. S. (1999). *Learning, working & playing in the digital age.* Retrieved December 31, 2003, from *http://serendip.brynmawr.edu/sci_edu/seelybrown/*

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and culture of learning. *Educational Researcher, 18*(1), 32-42.

Capurro, R. (2003). *Towards an information ecology.* Retrieved December 31, 2003, from *http://www.capurro.de/nordinf.htm#(9)*

Chan, T. W., Hue, C. W., Chou, C. Y., & Tzeng, J. L. (2001). Four spaces of network learning models. *Computers & Education, 37*, 141-161.

Cheng, K. M. (1997). Quality assurance in education: The East Asian perspective. In K. Watson, D. Modgil, & S. Modgil (Eds.), *Educational dilemmas: Debate and diversity: Vol. 4. Quality in education* (pp. 399-410). London: Cassell.

Collins, A., & Bielaczyc, K. (1997). Dreams of technology-supported learning communities. *Proceedings of the Sixth International Conference on Computer-Assisted Instruction,* Taipei, Taiwan.

Dillenbourg, P. (Ed.). (1999). *Collaborative learning: Cognitive and computational approaches.* Amsterdam: Pergamon.

Drucker, F. P. (1999). Knowledge worker productivity: The biggest challenge. *California Management Review, 41*(2), 79-94.

Gilbert, N. J., & Driscoll, M. P. (2002). Collaborative knowledge building: A case study. *Educational Technology Research and Development, 50*(1), 59-79.

Harasim, L. (Ed.). (1995). *Learning networks: A field guide to teaching and learning online.* Cambridge, MA: MIT Press.

Harmon, S. W., & Jones, M. G. (2001). An analysis of situated Web-based instruction. *Educational Media International, 38*(4), 271-279.

Hung, D. (2001). Theories of learning and computer-mediated instructional technologies. *Education Media International, 38*(4), 281-287.

Jonassen, D. H., Howland, J., Moore, J., & Marra, R. M. (2003). *Learning to solve problems with technology: A constructivist perspective* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice-Hall.

Jonassen, D. H., Peck, K. L., & Wilson, B. G. (1999). *Learning with technology: A constructivist perspective.* Upper Saddle River, NJ: Prentice Hall.

Krechevsky, M., & Stork, J. (2000). Challenging educational assumptions: Lessons from an Italian-American collaboration. *Cambridge Journal of Education, 30*(1), 57-74.

Lee, W. O. (1996). The cultural context for Chinese learners: Conceptions of learning in the Confucian tradition. In D. A. Watkins & J. B. Biggs (Eds.), *The Chinese learner: Cultural, psychological and contextual influences* (pp. 25-41). Hong Kong, China: Comparative Education Research Centre, The University of Hong Kong.

Lindberg, L. (2001). *Communities of learning: A new story of education for a new century.* Retrieved November 30, 2004, from *http://www.netdeva.com/learning*

Moursund, D. (1999). *Project-based learning using IT.* Eugene, OR: International Society for Technology in Education.

Rose, S. A., & Winterfeldt, H. F. (1998). Waking the sleeping giant: A learning community in social studies methods and technology. *Social Education, 62*(3), 151-152.

Scardamalia, M., & Bereiter, C. (1996). Engaging students in a knowledge society. *Educational Leadership, 54*(3), 6-10.

Scardamalia, M., & Bereiter, C. (2002). *Schools as knowledge building organizations.* Retrieved March 7, 2002, from *http://csile.oise.utoronto.ca/csile_biblio.html#ciar-understanding*

Scardamalia, M., Bereiter, C., & Lamon, M. (1995). The CSILE project: Trying to bring the classroom into world 3. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practices* (pp. 201-288). Cambridge, MA: Bradford Books/MIT Press.

Schütz, R. (2002, March 3). *Vygotsky and language acquisition.* Retrieved December 31, 2003, from *http://www.english.sk.com.br/sk-vygot.html*

Siemens, G. (2003). *Learning ecology, communities, and networks: Extending the classroom.* Retrieved October 17, 2003, from *http://www.elearnspace.org/Articles/learning_communities.htm*

Slevin, J. (2000). *The Internet and society.* Malden: Blackwell Publishers Ltd.

Van Der Stuyf, R. R. (2002, November 11). *Scaffolding as a teaching strategy.* Retrieved December 31, 2003, from *http://condor.admin.ccny.cuny.edu/~group4/Van%20Der%20Stuyf/Van%20Der%20Stuyf%20Paper.doc*

Vygotsky, L. S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Watkins, D. A., & Biggs, J. B. (Eds.). (2001). *Teaching the Chinese learner: Psychological and pedagogical perspectives* (2nd ed.). Hong Kong, China: Comparative Education Research Centre, The University of Hong Kong.

Wenger, E., McDermott, R., & Snyder, W. M. (2002). *Cultivating communities of practice: A guide to managing knowledge.* Boston: Harvard Business School Press.

Yuen, A. (2003). Fostering learning communities in classrooms: A case study of Hong Kong schools. *Education Media International, 40*(1/2), 153-162.

KEY TERMS

Anchored Learning Instructions: High learning efficiency with easier transferability of mental models and the facilitation of strategic problem-solving skills in ill-structured domains are emerged if instructions are anchored on a particular problem or set of problems.

CMC: Computer-mediated communication is defined as various uses of computer systems and networks for the transfer, storage, and retrieval of information among humans, allowing learning instructions to become more authentic and students to engage in collaborative project works in schools.

CMT: Computer-mediated technology points to the combination of technologies (e.g., hypermedia, handheld technologies, information networks, the

Internet, and other multimedia devices) that are utilized for computer-mediated communications.

Knowledge Building: In a knowledge-building environment, knowledge is brought into the environment and something is done collectively to it that enhances its value. The goal is to maximize the value added to knowledge: either the public knowledge represented in the community database or the private knowledge and skills of its individual learners. Knowledge building has three characteristics: (a) Knowledge building is not just a process, but it is aimed at creating a product; (b) its product is some kind of conceptual artifact, for instance, an explanation, design, historical account, or interpretation of a literacy work; and (c) a conceptual artifact is not something in the individual minds of the students and not something materialistic or visible but is nevertheless real, existing in the holistic works of student collaborative learning communities.

Learning Community: A collaborative learning community refers to a learning culture in which students are involved in a collective effort of understanding with an emphasis on diversity of expertise, shared objectives, learning how and why to learn, and sharing what is learned, thereby advancing the students' individual knowledge and sharing the community's knowledge.

Learning or Information Ecology: For preserving the chances of offering the complexity and potential plurality within the technological shaping of knowledge representation and diffusion, the learning- or information-ecology approach is indispensable for cultivating practical judgments concerning possible alternatives of action in a democratic society, providing the critical linguistic essences, and creating different historical kinds of cultural and technical information mixtures. Noteworthy is the fact that learning or knowledge involves a dynamic, living, and evolving state.

Metacognition: If students can develop metacognition, they can self-execute or self-govern their thinking processes, resulting in effective and efficient learning outcomes.

Project Learning or Project Works: Project learning is an iterative process of building knowledge, identifying important issues, solving problems, sharing results, discussing ideas, and making refinements. Through articulation, construction, collaboration, and reflection, students gain subject-specific knowledge and also enhance their metacognitive caliber.

Situated Learning: Situated learning is involved when learning instructions are offered in genuine living contexts with actual learning performance and effective learning outcomes.

Social Community: A common definition of social community has usually included three ingredients: (a) interpersonal networks that provide sociability, social support, and social capital to their members; (b) residence in a common locality, such as a village or neighborhood; and (c) solidarity sentiments and activities.

ZPD: The zone of proximal development is the difference between the child's capacity to solve problems on his or her own, and his or her capacity to solve them with the assistance of someone else.

Constructing a Globalized E-Commerce Site

Tom S. Chan
Southern New Hampshire University, USA

INTRODUCTION

Traditional boundaries and marketplace definitions are fast becoming irrelevant due to globalization. According to recent statistics, there are approximately 208 million English speakers and 608 million non-English speakers online, and 64.2% of Web users speak a native language other than English (Global Reach, 2004). The world outside of English-speaking countries is obviously coming online fast. As with activities such as TV, radio and print, people surf in their own language. A single-language Web site simply could not provide good visibility and accessibility in this age of globalize Internet. In this article, we will focus on the approaches in the construction of an effective globalized e-commerce Web site.

A SHORT TOUR OF E-COMMERCE SITES

The 1990s was a period of spectacular growth in the United States (U.S.). The commercialization of the Internet spawned a new type of company without a storefront and who existed only in cyberspace. They became the darlings of the new economy, and traditional brick-and-mortar retailers have been scoffed off as part of the old economy. Of course, this irrational exuberance is hampered with a heavy dose of reality with the dot.com bust in 2000. Yet, the trends initiated by the dot.com start-ups; that is, conducting commerce electronically, are mimicked by traditional businesses. The Internet is a haven, and imperative for commerce. And, not only for business-to-consumer transactions; business-to-business applications are also becoming more popular.

While there are endless possibilities for products and services on the Internet, e-commerce sites can be classified into a few broad categories: brochure, content, account and transaction sites. Both brochure and content sites provide useful information for customers. A brochure site is an electronic version of a printed brochure. It provides information about the company and its products and services, where contents tend to be very static. A content site generates revenue by selling advertisement on the site. It attracts and maintains traffic by offering unique information, and content must be dynamic and updated regularly. An account site allows customers to manage their account; for example, make address changes. A transaction site enables customers to conduct business transactions; for example, ordering a product. Unlike brochure and content sites, security safeguards such as password validation and data encryption are mandatory. Typical e-commerce sites today are multidimensional. For example, a mutual fund company's site provides company information and current market news, but it also allows customers to change account information and sell and buy funds.

A STANDARD SITE CONSTRUCTION METHODOLOGY

Over the past decade, e-commerce site development methodology has become standardized following the model of system development life cycle, with activities including planning, analysis, design, implementation and support. Launching a business on the Internet requires careful planning and strategizing. Planning requires coming up with a vision and a business plan, defining target audiences and setting both short- and long-range goals. Analysis means defining requirements and exploring and understanding the problem domain to determine the site's purpose and functionality. Design requires selecting hardware and software, and determining site structure, navigation, layout and security issues. Implementation means building the site and placing it on the Internet. Support requires maintaining the site, supporting its customers and conducting periodic upgrades to improve its performance and usability.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

A successful globalize e-commerce site must strike a balance between the need for uniformity and accommodating variations. While most contents are identical (though they may be presented in different languages), some content inevitably varies and only is relevant for the locals. A site with a globalize reach must be adapted for both localization and optimization. While there are many issues to consider in the construction of an e-commerce site, our primary focus here deals with aspects particularly relevant to a globalize site, including issues of site specification, customer research and branding, site structure, navigation and search engine optimization.

Site Specification and Functionality

It is very easy to confuse an e-commerce site with the corporation. An inescapable truth, the corporation owns its Web site. The corporation also handles legal, marketing, public relationships, human resources and many other matters associated with running a business. It is important to understand that the site serves a business, and not the other way around (Miletsky, 2002). All corporations have a mission statement and associated strategies. A site exists to serve the corporation, and the site's functionality should reflect this reality. Therefore, it is important to ask: How could the site help the corporation in the execution of its business plan?

Globalization may increase a corporation's market and nurture opportunities, but it may not be for everyone. What role is the globalize e-commerce site playing? Could the corporation's product or service have market potential in other countries? While globalization creates new opportunities, it also invites new competition. A corporation naturally has an idea of local competitors. Before globalizing, know the competitors in the international market. Because competition may vary from country to country, functions and priorities of the site need to adjust accordingly. For the same corporation and same product, approaches may need to vary for different localities. Given the inevitability of globalization, internationalizing the corporate e-commerce site may be a necessary defensive against competitors making inroads into one's market.

Understand One's Customers

With hundreds of thousands of e-commerce sites to visit, customers are faced with more choices than ever. The vast amount of information found on the Internet is daunting, not only to the shoppers but to corporations, as well. What do users want when they come to the site? Are they individual consumers, commercial businesses or government agencies? The clients access the site for information. But, do they perform sales online, and what about exchanges and returns? How are they able to perform these functions from the site? What are the implications for multicultural users? What technologies will be needed on the site to support these functions?

The cardinal rule in building a functional site is to understand its users. However, the same site operating in different countries may target different audiences. For example, teenagers in the U.S. may have more discretionary spending than their European or Asian counterparts. A sport sneaker site targeting teenagers in the U.S. may target adults in another country. The per capita income of a particular region will definitely affect sales potential and project feasibility. On the other hand, the same target in a different country may also have different needs. Site planners must consider the daily functions and preferences of the local customers and organize sites to support those functions (Red, 2002). For example, while Web sites are becoming major portals for information distribution, many Asian cultures value personal contact. It is important to provide contact phone numbers instead of business downloads for a personal touch, or it could be viewed as rude and impolite.

A large part of building a successful e-commerce site is being aware of the ways in which customers reach you, which pages are most popular, and what sticky features are most effective. Partnerships with other e-businesses can also help attract new customers. Local sites in different countries also need to be accessible to each other. How will the partner and local sites be linked together? For customers unfamiliar with the site, will they know what the corporation offers, and be able to find the things they need? When developing a local site, the original home site can only be used as a reference. The site must localize properly. A key activity is to inventory all information local customers need and want to access. Next, consider

how they will move among the different types of information. Since the Web is not a linear vehicle, one should explore the many ways people might want to come in contact with the content in one's site. This will help shape how content should be organized.

Branding for Consistency

Prior to venturing into site design, we must first discuss branding. A corporation's identity speaks volumes. It lives and breathes in every place and every way the organization presents itself. A familiar and successful brand requires years of careful cultivation. Strong brands persist, and early presence in a field is a prime factor in strong brand establishment. Apart from being a relatively new venue, branding in e-commerce is critical, because customers have so much choice. They are bombarded by advertisements, and competitors are only one click away. In an environment of sensory overload, customers are far more dependent on brand loyalty when shopping and conducting business on the Internet (Tsiames & Siomkos, 2003).

A corporation, especially one with global reaches, must project an effective and consistent identity across all its Web sites and pages. Consistently maintained, the pages present a unified corporate image across markets and geographic regions, identifying different businesses that are part of the same organization, reinforcing the corporation's collective attributes and the implied commitments to its employees, customers, investors and other stakeholders. Branding is particularly important for multinational corporations because of their vast size, geographic separation and local autonomy. The success or failure of site branding depends entirely on the effectiveness and uniformity of the organization, linkage of pages and presentation of its contents. Inevitably, there will always be creative tensions between uniformity imposed by the global mission and brand vs. adaptation towards local customers and competitors. Always aim at obtaining an effective balance between these two requirements.

Figure 1. Structure of a bilingual Web site

- abcCorp.com
  - siteIndex_en.htm
  - siteIndex_zh.htm
  - <en>
    - aboutUs.htm
    - <product>
    - <service>
  - <zh>
    - aboutUs.htm
    - <product>
    - <service>
  - <share>
    - logo.jpg
    - banner.gif

Structuring a Multilingual Site

In general, it is not a good practice to intermix different language scripts on the same document, for aesthetic reasons. While some of us are bilingual, it is a rarity to be multilingual. A multilingual page means most customers will not be able to understand a large portion of the display. They will either get confused or annoyed quickly. Therefore, it is best to structure a multilingual site using mirror pages of different languages.

A Web site is a tree model, with each leaf representing a document or a branch linked to a set of documents. The home page forms the root index to each document or branch. A classic multilingual site structure contains indexes in various languages, branches for each language and a common directory. Figure 1 shows an example of a bilingual site that supports English and Chinese. The index of the main language (English) and mirrored indexes in a different language (Chinese) are stored in the root directory. Subdirectories "en" (English) and "zh" (Chinese) contain actual site contents and are identically structured, with each branch containing an introduction and branches for products and services. Naturally, there is also a "share" subdirectory for common files, such as corporate logo and background images (Texin & Savourel, 2002).

Files and directories in different languages can use a suffix or prefix as an identifier. However, multilingual standards for the Web are well established today; identifier selection can no longer be ad-lib. There should be a two-letter language code (ISO, 2002), followed by a two-letter country subcode (ISO, 2004) when needed, as defined by the International Organization for Standardization.

Naming and Localization

Directory and file names should be meaningful and consistent; translating names to different languages should be avoided. Under the structure in Figure 1, except the indexes, different names for the same file in different languages are not required. An index page would contain links—for example, a pull-down menu—to access an index of other languages. A major design requirement for any site is structure clarity and localization. This structure maximizes relative URLs, minimizing numbers and changes of links while still providing emphasis and localization for the individual language. It also allows for search engine optimization, a matter that will be discussed later.

A common technique in managing a multilingual site is to use cookies to remember language preferences and then generate the URL dynamically. With page names standardized, we can resolve links at run time. For example, a Chinese page has a suffix "_zh." By examining the browser's cookie, we know the customer visited the Chinese page in the last visit. Thus, we should forward the customer to the Chinese index page. We can make a script to append _zh to file name siteIndex to generate the destination URL. If, on the other hand, while the customer is reading the product page in Chinese he or she wishes to visit its English version, the URL can also be generated dynamically by substituting "/zh/" with "/en/" in the current URL string. Naturally, it would be easier with a multiple address schema that each language has its own domain name; for example, "www.abc.com" for the English site and "www.abc.zh" for the Chinese site. Unfortunately, multiple domain names involve extra cost, both in development and administration. As a practical matter, most sites have only one domain address (Chan, 2003).

Navigation for Inconsistency

The design for a welcome page is very tricky for a multilingual Web site. While each supported language has its own index, where should be a customer be redirected if we cannot determine a preference? A common design is to splash a greeting page and prompt customers for a selection. Since the consumer's language is still undetermined, the page typically contains only images with minimal to no textual description, leading to a home page without a single word, just the corporate logo and image buttons. Not only is this design awkward from an aesthetic angle, it also leaves no content for a search engine to categorize. It is far better to identify a language group with the largest users, make that index the default home page, and provide links from that page to an index page of the other supported languages.

Consistency is a crucial aspect in navigation design. With a consistent structure, customers would not get confused while surfing the site (Lynch & Horton, 2001). Some sites provide mirror contents of its pages in different languages. A common navigational feature would be icons either in graphics, such as national flags, or foreign scripts. Customers can click on the link to view contents in another language. In a multilingual site, consistency becomes problematic because of content differences. For example, some contents available in English may have no meaning or do not apply to Chinese customers. Even when contents are identical, translation may not produce pages in a neat one-to-one map. A customer who comes from English pages to a Chinese page will likely be confused, encountering a more or less different structure. From a functional perspective, accessing mirrored content has very little utility. After all, except for academics or recreation, why would an English customer reading product descriptions in English want to read its counterpart in Chinese, or vise versa? Besides the index page, links to the mirrored content of another language should be discouraged, and consistent navigation should enforce only inside, not outside, a single language hierarchy structure.

Optimizing for Spiders

A necessary step for visibility is to submit the site to search engines. A search engine has three major components: spider, index and algorithm (Sullivan, 2004). The spider visits the site's URL submitted by the corporation, reads it and follows links to other pages, generating an index to report to user queries. While search engine's algorithm is a trade secret, the main rule involves location and frequency. The spider checks for keywords that appear in headers and the home page. The further away the link, the less search engines consider its importance. They assume that relevant words to the site will be mentioned close to the beginning. A search engine also analyzes how

often keywords appear. Those with a higher frequency are deemed more relevant. If spiders do a poor job reading the site, identified keywords will not reflect its products and services. The site will probably not get visitors who are potential customers, or it may not get many hits at all.

A greeting splash page without text but with logo and buttons would, therefore, make a very poor choice as a home page for search engine submission. First, there are no keywords for the spider to follow. Second, relevant contents become one additional link away. Third, the spider would be unable to read most of the links and contents, as they are mostly in foreign scripts. Since leading search engines in each country is a local engine and indexing sites in the local language only, one should submit an index page of the local language and register it as a local dot com or, where possible, as a purely local domain. Unfortunately, as stated earlier, multiple domain names are expensive; a local dot com or local domain is rarely a practical alternative (Chan, 2004).

To optimize the site for search engines, concentrate mostly on the content directly linked to the home page—both internal and external—by working important keywords in both the content and link text as much as possible. Follow that up to a lesser extent with internal pages a few links away. The file name and page title should contain keywords, as most search engines look to them as an indication of content. Spamming a search engine is when a keyword is repeated over and over in an attempt to gain better relevance. Such practice should never be considered. If a site is considered spam, it may be banned from the search engine for life or, at the very least, ranking will be severely penalized. More information is available from the Notess.com (2004) Web site regarding the particular characteristics of popular search engines.

FUTURE TRENDS

In the global economy, increasing numbers of companies need their computing systems to support multiple languages. With the Windows XP release, Microsoft makes available 24 localized versions of Windows in addition to English (Microsoft, 2001). Users can display, input, edit and print documents in hundreds of languages. At the same time, the Internet is internationalizing, and its standards are modernizing. While the original Web is designed around the ISO Latin-1 character set, the modern system uses UTF-8, a standard designed for all computer platforms and all languages (Unicode, 2003). The HTML specification has also been extended to support globalize character set and multilingual contents (W3C, 1999). As more computer platforms and applications are configured to support local languages, proper adherence to a multilingual Web standard will be mandatory, even when building U.S.-only sites.

CONCLUSION

A successful globalize site design involves more than translating content from one language to another. It requires proper localization of requirement definition and internationalization of the site design for effective structure, navigation and indexing. As global exchanges become a common practice, proper implementation of a multilingual Web structure and standard is crucial for any e-commerce site. To that end, most operating systems, applications, Web editors and browsers today are configurable to support and construct Web sites that meet international standards. As the Internet becomes globalized and Web sites continue to be the major portal for interfacing with customers, a site constructed properly will empower an organization to reach audiences all over the world as easily as if they are living next door.

REFERENCES

Chan, T. (2003). *Building multilingual Web sites for today's global network.* Paper presented at E-Learning World Conference.

Chan, T. (2004). *Web site design for optimal global visibility.* Paper presented at International Academy of Business Disciplines World Conference.

Global Reach (2004). Global Internet Statistics. Retrieved June 2004 from *global-reach.biz/globstats*

International Organization for Standardization. (2002). ISO 639 – Codes for the representation of names of languages.

International Organization for Standardization. (2004). ISO 3166 – 1, 2 & 3. Codes for the representation of names of countries and their subdivisions.

Lynch, P., & Horton, S. (2001). *Web style guide: Basic design principles for creating Web sites.* Yale University Press.

Microsoft. (2001). Windows XP Professional overview, multilingual support. Retrieved June 2004 from *www.microsoft.com/windowsxp/pro/evaluation/overviews/multilingual.asp*

Notess, G. (2004). Search engine features chart. Retrieved June 2004 from *searchengineshowdown.com/features/*

Red, K.B. (2002). *Considerations for connecting with a global audience.* Paper presented at International WWW Conference.

Sullivan, D. (2004). Optimizing for crawlers. Retrieved June 2004 from *www.searchenginewatch.com/Webmasters/article.php/2167921*

Texin, T., & Savourel, Y. (2002). *Web internationalization standard and practices*. Paper presented at International Unicode Conference.

Tsiames, I., & Siomkos, G. (2003). E-brands: The decisive factors in creating a winning brand in the net. *Journal of Internet Marketing, 4*(1).

Unicode Consortium, The. (2003). The Unicode standard, version 4.0. Retrieved June 2004 from *www.unicode.org/versions/Unicode4.0.0/*

World Wide Web Consortium. (1999). The HTML 4.01 Specification. Retrieved June 2004 from *www.w3.org/TR/html401/*

C

KEY TERMS

Brand: The promise that a Web site, company, product or service makes to its customers.

E-Commerce: Conducting business and financial transactions online via electronic means.

Globalize: Business issues associated with taking a product global. It involves both internationalization and localization.

Internationalize: Generalizing a design so that it can handle multiple languages content.

Localize: Design linguistically and culturally appropriate to the locality where a product or service is being used and sold.

Search Engine: A program that indexes Web documents, then attempts to match documents relevant to a user's query requests.

Site Specification: A design document for a Web site specifying its objectives and functionality.

UTF-8: Unicode Transformation Format 8 bits; the byte-oriented encoding form of Unicode.

Visibility: Points and quality of presence on where potential customers can find a Web site.

Consumer Attitude in Electronic Commerce

Yuan Gao
Ramapo College of New Jersey, USA

INTRODUCTION

As a valuable communications medium, the World Wide Web has undoubtedly become an important playground of commercial activities. Founded on a hypermedia document system, this medium plays a critical role in getting messages across to visitors, who may be current or perspective customers. In business-to-consumer (B2C) Web sites, companies are engaged in a wide range of activities including marketing, advertising, promotion, sales, and customer service and support (Berthon, Pitt, & Watson, 1996; Singh & Dalal, 1999). As a result, practitioners and scholars alike have started to examine various techniques ranging from the overall structure of the online retailing interface to individual features as banners, animation, sound, video, interstitials, and popup ads (Rodgers & Thorson, 2000; Westland & Au, 1998). Consumers are the ultimate judges of the success of any online retailing site, and consumer perceptions mediate content factors in influencing their attitude toward electronic commerce as well as individual e-tailing sites, complementing the roles played by Web site content in shaping consumer attitude.

BACKGROUND

In traditional advertising research, Olney et al. (1991) outlined a chain of links where both content and form variables were examined as predictors of attention, memory, recall, click-through, informativeness, attractiveness, and attitude. An evaluation of these outcome variables in the Web context necessarily involves new dimensions that require a higher degree of comprehensiveness due to the volume and scope of a Web site in comparison to print or TV ads. For example, Rogers and Thorson (2000) argue for the consideration in interactive marketing of such techniques as banners, sponsorships, interstitials, popup windows, and hyperlinks over and beyond ad features found in traditional media, such as color, size, and typeface in the print media, and audio, sound level, animation, and movement in broadcast.

Factors related to consumer behavior, attitude, and perceptions in the online environment have been examined in recent research (Chen & Wells, 1999; Coyle & Thorson, 2001; Ducoffe, 1996; Eighmey, 1997; Gao, Koufaris, & Ducoffe, 2004; Koufaris, 2002; Koufaris, Kambil, & Labarbera, 2001; Vijayasarathy, 2003). Consumer attitude mediates the effect of systems characteristics on behavioral intentions such as intention to revisit and intention to purchase products from the sponsoring companies. Past research has shown that the value of advertising derives from informative claims in an entertaining form (Ducoffe, 1995), while Web site users similarly appreciate information in an enjoyable context (Eighmey, 1997). Koufaris et al. (2001) found shopping enjoyment a significant factor attracting return visits. We consider information, entertainment, and site organization major measurement criteria and perceptual antecedents that affect user attitude toward communications messages presented through the Web (Ducoffe, 1996) and attitude toward the Web site as a whole (Chen & Wells, 1999).

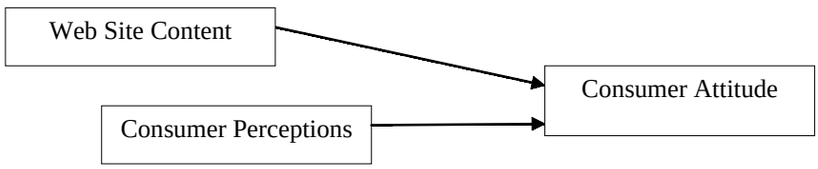
This article provides an overview of current research on factors influencing consumer attitude and related behavioral consequences in electronic commerce. It reviews and synthesizes research from two perspectives: Web site content and consumer perceptions. The next section discusses research uncovering content factors that impact consumer attitude or other attitudinal consequences, while the following section examines consumers' perceptual dimensions that influence their attitude in Web-based commerce. The following diagram serves as a schema in guiding the presentation of our framework.

WEB SITE CONTENT

Content is king (Nielsen, 1999, 2003). Message content believed to be informative by a marketer

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Schema of factors influencing consumer attitude in electronic commerce



needs to be substantiated by consumer feedback. In analyzing the informativeness of a message, content analysis complements attitudinal research by pointing out the types of information and Web site features that make a site informative, entertaining, or irritating. Web site content discussed in this article contains information, presentation attributes, and system design features.

Information

In traditional advertising research, Resnik and Stern (1977) developed a content analysis method through codifying each advertising message via 14 evaluative cues. Numerous studies used this procedure in analyzing ad messages in various media, including magazine, TV, and newspaper advertising (Abernethy & Franke, 1996). Among those studies, a few tried to connect message content with informativeness. For example, Soley and Reid (1983) find that quality, components or content, price or value, and availability information affected perceived informativeness, while the quantity of information did not. Ylikoski (1994) finds moderate support for the connection between the amount of informative claims and perceived informativeness in an experimental study involving automobile advertisements.

In a similar approach, Aaker and Norris (1982) developed a list of 20 characteristic descriptors intended to explain a commercial message's informativeness. They find hard sell versus soft sell, product class orientation, and number of distinct claims, e.g., on product quality or performance, are the most significant predictors of informativeness from a study based on 524 TV commercials.

Adapted versions of the content analysis method have been applied to analyzing Web advertising and Web sites (Ghose & Dou, 1998; Philport & Arbittier, 1997). Other studies have attempted to categorize Web site content based on technology features (Huizingh, 2000; Palmer & Griffith, 1998). The development of these approaches demonstrates the complexity of Web-based communications and reflects a need to have a more sophisticated method to understand what constitute an effective Web site. Thus, we must inevitably turn our attention to design features and techniques that contribute to the delivery of entertainment, in addition to information, in this new medium.

Presentation Attitudes

Philport and Arbittier (1997) studied content from over 2000 commercial communications messages across three established media, that is, TV, magazines, and newspapers, along with that on the Internet. The adoption of variables such as product demonstration or display, special effect techniques like fantasy, and the employment of humor reflects an attempt by researchers to assess message appeal enhanced by entertaining features. Philport and Arbittier (1997) find no distinguishing characteristic of banner ads from other media ads. Their study suggests that the impact of a message delivered through a banner is fairly limited, and the integral collection of hypermedia-based documents, related image files, and system functions as a whole is a better candidate for examining the effectiveness of Web-based communications.

Ghose and Dou (1998) linked the number of content attributes with site appeal measured by being listed in Lycos top 5% of Web sites and found that a greater degree of interactivity and more available online entertainment features increase site appeal. Huizingh (2000) content-analyzed 651 companies from Yahoo and Dutch Yellow Pages using a battery including elements like pictures, jokes, cartoons, games, and video clips. He found that entertainment

features appear in about one-third of all sites, and that larger sites tend to be more entertaining.

Existing literature has also touched upon the effect of media formats on consumer attitude, especially in interactive marketing research (Bezjian-Avery, Calder, & Iacobucci, 1998; Coyle & Thorson, 2001; Rodger & Thorson, 2000). Bezjian-Avery et al. (1998) tested the impact of visual and nonlinear information presentation on consumer attitude toward products. Steuer (1992) provides a theoretical discussion on the mediating impact of communications technology on a person's perception of his/her environment, termed telepresence, determined by the three dimensions of interactivity, including speed, range, and mapping, and the two dimensions of vividness, including breadth and depth. Coyle and Thorson (2001) associated interactivity and vividness with perceived telepresence and consumer attitude toward brand and site, and find that both perceived interactivity and perceived vividness contribute to attitude toward the site and subsequent intention to return to a site.

Alongside entertainment and information, Chen and Wells (1999) also identify a factor "organization" that describes the structure or navigational ease of a site. Eighmey (1997) finds that structure and design of a Web site are important factors contributing to better perceptions of Web sites. System design features may enhance visitor experience and efficiency in information retrieval, and thus contribute to both perceived informativeness and reduced irritation. The following are some recent studies examining the effects of system design feature in e-commerce sites.

System Design Features

Relating to site features, Lohse and Spiller (1998) performed a study measuring 32 user interface features at 28 online retail stores against store traffic and sales. They conclude that online store traffic and sales are influenced by customer interfaces. In particular, they found that an FAQ page, promotional activities, and better organization of the product menu have significant positive influences on traffic and sales. Huizingh (2000) considers the complexity of the navigation structure and search function design features and finds that more complex structures are found in larger Web sites, which are also more likely to have a search mechanism. Recognizing content the most important element of a Web site, Nielsen (1997, 1999, 2000) points out a few critical areas of design that determine the success or failure of a Web site: speed, quality of a search mechanism, and clarity of structure and navigation.

Research addressing the impact of different digital retailing interfaces (Westland & Au, 1999) reveals that virtual reality storefronts increase a consumer's time spent searching for products but do not significantly increase sales. In the field of human-computer interaction, significant research has been done relating network quality of service with usability and user satisfaction. One such factor affecting quality of service is system speed. The effect of system speed on user reactions was studied in both the traditional and Web-based computing environments (Sears & Jacko, 2000). Nielsen (1997) argued, based on a combination of human factors and computer networking, "speed must be the overriding design criterion." He asserts that research has shown that users need a response time of less than one second, moving from one page to another, based on traditional research in human factors.

In a study linking the use of interruption implemented via pop-up windows, Xia and Sudharshan (2000) manipulated the frequency of interruptions and found that interruptions had a negative impact on consumer shopping experiences. Intrusive formats of advertising like interstitials are found to have "backlash risks" in this new medium (Johnson, Slack, & Keane, 1999). Gao et al. (2004) find that continuously running animation and unexpected popup ads have a positive association with perceived irritation, and contribute negatively to attitude toward the site.

To summarize, along with information content and presentation attributes, system design features are some of the applications of current information technology that may influence consumer perceptions of Web sites and their attitude toward those Web sites.

CONSUMER PERCEPTIONS AND ATTITUDE

Informativeness is a perception (Ducoffe, 1995; Hunt, 1976). Research in marketing and advertising has also focused on consumer perceptions of a communications message, and how these percep-

tions influence advertising value and consumer attitude (Chen & Wells, 1999; Ducoffe, 1995, 1996). Information contained in a commercial message is believed to be individual-specific and cannot be measured objectively (Hunt, 1976). Content analysis researchers seem to concur on these points. Resnik and Stern, being pioneers in applying content analysis to advertising message content, acknowledge that it would be unrealistic to create an infallible instrument to measure information because information is in the eye of the beholder (1977). However, they maintain that without concrete information for intelligent comparisons, consumers may not be able to make efficient purchase decisions (Stern & Resnik, 1991).

Perceived informativeness, entertainment, and irritation have been shown to affect consumer attitude toward Web advertising, considered by 57% of respondents in one study to include a firm's entire Web site (Ducoffe, 1996).

An online shopper's experience with an e-commerce site is similar to exposure to advertising. The shopper's assessment of the value of advertising can be drawn from exchange theory. An exchange is a relationship that involves continuous actions and reactions between two parties until one of the parties distances itself from such a relationship when it sees it as no longer appropriate (Houston & Gassenheimer, 1987). The value derived from such an exchange from the consumer's perspective is an important factor in further engagement of the consumer in this relationship. Advertising value is "a subjective evaluation of the relative worth or utility of advertising to consumers" (Ducoffe, 1995, p.1). Such a definition is consistent with a generic definition formulated by Zeithaml (1988), who defined value of an exchange to be "the consumer's overall assessment of the utility of a product based on perceptions of what is received and what is given" (p.14).

A visit to a Web site is a form of exchange in which the visitor spends time learning information from, and perhaps enjoying entertainment at the site. In order for such a relationship to sustain itself, the benefits must outweigh the costs. Considering information and entertainment two major benefits a consumer derives from visiting a commercial site, a Web site's value is enhanced by more informative and entertaining presentations of products and services.

However, the value of a Web site, like advertising, is individual specific. One consumer may find what she needs at a site and perceive the site high in value, while another person may find it low in value because of the lack of information he wants. Someone may find a site high in value because it fulfills his entertainment needs while another person may not.

Relating to measures of general likability of an advertisement, attitude toward the ad (Aad) has been found to have both cognitive and affective antecedents, where deliberate, effortful, and centrally processed evaluations result in said cognitive dimensions (Brown & Stayman, 1992; Ducoffe, 1995; Muehling & McCann, 1993). MacKenzie and Lutz (1989) argue that such evaluations can be viewed as antecedents to consumer attitude toward an advertisement. Attitude toward the site (Ast) is a measure parallel to attitude toward the Ad (Aad) and was developed in response to a need to evaluate site effectiveness, like using Aad to evaluate advertising in traditional media (Chen & Wells, 1999). Aad has been considered a mediator of advertising response (Shimp, 1981). Since Aad has been found to influence brand attitudes and purchase intentions (Brown & Stayman, 1992), it is considered an important factor for marketing and advertising strategies. Attitude toward the site is considered an equally useful indicator of "Web users' predispositions to respond favorably or unfavorably to Web content in natural exposure situations" (Chen & Wells, 1999). They find that 55% of variance in attitude toward a Web site are explained by entertainment, informativeness, and organization factors. Eighmey (1997) finds that entertainment value, amount of information and its accessibility, and approach used in site presentation, account for over 50% of the variance in user perceptions of Web site effectiveness.

Ducoffe (1995, 1996) finds a significant positive .65 correlation between informativeness and advertising value in traditional media and .73 correlation in Web advertising, and a significant positive .48 correlation between entertainment and advertising value in traditional media and .76 correlation in Web advertising. Chen and Wells (1999) find a positive correlation of .68 between informativeness and attitude toward a site, and a positive .51 correlation between entertainment and attitude toward a site. Ducoffe (1995, 1996) finds a significant and negative correlation of -.52 between irritation and advertising value in traditional media and -.57 in Web advertising. We maintain that perceived disorganization is one major factor contrib-

uting to perceived irritation. Chen and Wells (1999) find a positive .44 correlation between "organization" and attitude toward a site.

In summary, from the perspective of consumer perceptions, we consider the perception of a Web site being informative, entertaining, and organized as three major antecedents positively associated with consumer attitude in e-commerce.

FUTURE TRENDS

We summarize our discussion in the previous two sections into the following general diagram. This diagram provides a framework for further thinking in the development of e-commerce systems in general and systems design influencing consumer attitude in particular. In this diagram, we recognize that Web site content may influence both a consumer's perception and his or her attitude, thus Web site content features could have both a direct and indirect impact on consumer attitude. Nonetheless, the perceptual dimensions capture a much broader realm of variables and explain a larger percentage of variance in attitude than those by individual features and content, especially in behavioral science research (Cohen, 1988).

Internet technology and e-commerce continue to grow. How to achieve a competitive advantage through utilizing the advancement in information technology to support a firm's product offerings is a question faced by many e-commerce firms. In accordance with our review of literature, we suggest that both marketing executives and system developers of e-commerce Web sites pay attention to the underlying connectivity between system design and consumer behavior, and strive to closely examine the issue of integrating technological characteristics and marketing communications in the Web context. We offer the following guidelines.

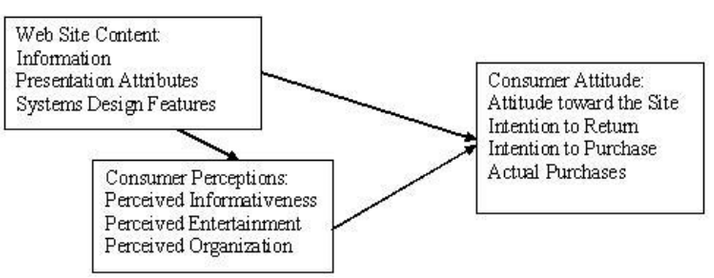
First, online shoppers value information that is essential to their purchase decisions. It has been demonstrated that consumers value commercial messages that deliver "the most informative claims an advertiser is capable of delivering" in a most entertaining form (Ducoffe, 1996).

Second, consumers appreciate entertainment. An enjoyable experience increases customer retention and loyalty (Koufaris et al., 2001). An entertaining Web site helps to retain not only repeat visitors, but also chance surfers. It is imperative that Web site developers make customer experience their first priority when incorporating features and attributes into a Web site.

Third, consumers' attitude is enhanced by product experience that is more direct than simple text and images. Direct experience, virtual reality, and telepresence help deliver a message in a more informative and entertaining way.

Last but not least, Web sites should be cautious when using intrusive means of message delivery such as popup ads and animation. Using pop-up ads to push information to the consumer is sometimes a viable technique. Johnson, Slack, and Keane (1999) found that 69% surveyed consider pop-up ads annoying and 23% would not return to that site. Visitors to a Web site do not like interruptions, even those containing information closely related to products sold at the site (Gao et al., 2004). Such techniques should be reserved for mission-critical messages that otherwise cannot be effectively deployed.

Figure 2. Factors influencing consumer attitude in electronic commerce



CONCLUSION

The study of consumer attitude in e-commerce has not been widely explored, and the true effectiveness of any presentation attribute awaits further examination. We maintain that presentation attributes communicate much non-product information that can affect company image and visitor attitude toward products and the site. As a relatively new communications medium, the Internet provides message creators added flexibility and functionality in message delivery. Marketers can take advantage of the opportunities of incorporating system designs that further enhance a visitor's experience while visiting a Web site. Attitude is an affection. Future research should also explore the connection between presentation attributes and consumer perceptions, because the connection between what a system designer puts into a Web site and how an online visitor perceives it is the focal point where the interests of the marketers and the consumers meet.

REFERENCES

Aaker, D.A. & Norris, D. (1982). Characteristics of TV commercials perceived as informative. *Journal of Advertising Research*, *22*(2), 61-70.

Abernethy, A.M. & Franke, G.R. (1996). The information content of advertising: a meta-analysis. *Journal of Advertising*, *15*(2), 1-17.

Berthon, P., Pitt, L.F., & Watson, R.T. (1996). The World Wide Web as an advertising medium: Toward an understanding of conversion efficiency. *Journal of Advertising Research, 36*(1), 43-54.

Bezjian-Avery, A., Calder, B., & Iacobucci, D. (1998). New media interactive advertising vs. traditional advertising. *Journal of Advertising Research*, *38*(4), 23-32.

Brown, S.P. & Stayman, D.M. (1992). Antecedents and consequences of attitude toward the ad: A meta -analysis. *Journal of Consumer Research, 19*(1), 34-51.

Chen, Q. & Wells, W.D. (1999). Attitude toward the site. *Journal of Advertising Research*, *39*(5), 27-38.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coyle, J.R., & Thorson, E. (2001). The effects of progressive levels of interactivity and vividness in Web marketing sites. *Journal of Advertising*, *30*(3), 65-77.

Ducoffe, R.H. (1995). How consumers assess the value of advertising. *Journal of Current Issues and Research in Advertising, 17*(1), 1-18.

Ducoffe, R.H. (1996). Advertising value and advertising on the Web. *Journal of Advertising Research*, *36*(5), 21-34.

Eighmey, J. (1997). Profiling user responses to commercial Web site. *Journal of Advertising Research*, *37*(3), 59-66.

Gao, Y., Koufaris, M., & Ducoffe, R. (2004). An experimental study of the effects of promotional techniques in Web-based commerce. *Journal of Electronic Commerce in Organizations*, *2*(3), 1-21.

Ghose, S. & Dou, W. (1998). Interactive functions and their impact on the appeal of the Internet presence sites. *Journal of Advertising Research, 38*(2), 29-43.

Houston, F.S. & Gassenheimer, J.B. (1987). Marketing and exchange. *Journal of Marketing*, *51*(4), 3-18.

Huizingh, E.K.R.E. (2000). The content and design of Web sites: An empirical study. *Information & Management, 37*, 123-134.

Hunt, S.D. (1976). The nature and scope of marketing. *Journal of Marketing*, *44*(3), 17-28.

Johnson, M., Slack, M., & Keane, P. (1999). *Inside the mind of the online consumer — Increasing advertising effectiveness.* Jupiter Research. Accessed August 19, 1999, at *www.jupiter.com*

Koufaris, M. (2002). Applying the technology acceptance model and flow theory to online consumer behavior. *Information Systems Research*, *13*(2), 205-223.

Koufaris, M., Kambil, M.A., & Labarbera, P.A. (2001). Consumer behavior in Web-based com-

merce: An empirical study. *International Journal of Electronic Commerce*, *6*(2), 131-154.

Lohse, G.L. & Spiller, P. (1998). Electronic shopping. *Communications of the ACM*, *41*(7), 81-86.

MacKenzie, S.B. & Lutz, R.J. (1989) An empirical examination of the structural antecedents of attitude toward the ad in an advertising pretesting context. *Journal of Marketing*, *53*(2), 48-65.

Muehling, D.D. & McCann, M. (1993). Attitude toward the ad: A review. *Journal of Current Issues and Research in Advertising*, *15*(2), 25-58.

Nielsen, J. (1996). Top ten mistakes in Web design, *Jakob Nielsen's Alertbox*, May, 1996, accessed at *www.useit.com/alertbox/*

Nielsen, J. (1997). The need for speed. *Jakob Nielsen's Alertbox*, March, 1997, accessed at *www.useit.com/alertbox/*

Nielsen, J. (1999). User interface directions for the Web. *Communications of the ACM*, *42*(1), 65-72.

Nielsen, J. (2000). Is navigation useful? *Jakob Nielsen's Alertbox*, January, 2000 accessed at *www.useit.com/alertbox/*

Nielsen, J. (2003). Making Web advertisements work. *Jakob Nielsen's Alertbox*, May, 2003, accessed at *www.useit.com/alertbox/*

Palmer, J.W. & Griffith, D.A. (1998). An emerging model of Web site design for marketing. *Communications of the ACM*, *41*(3), 45-51.

Philport, J.C. & Arbittier, J. (1997). Advertising: Brand communications styles in established media and the Internet. *Journal of Advertising Research*, *37*(2), 68-76.

Resnik, A. & Stern, B.L. (1977). An analysis of information content in television advertising. *Journal of Marketing*, *41*(1), 50-53.

Rodgers, S., & Thorson, E. (2000). The interactive advertising model: How users perceive and process online ads. *Journal of Interactive Advertising, 1*(1). Accessed at *jiad.org/*

Sears, A. & Jacko, J.A. (2000). Understanding the relation between network quality of service and the usability of distributed multimedia documents. *Human-Computer Interaction*, *15*, 43-68.

Shimp, T.A. (1981). Attitude toward the ad as a mediator of consumer brand choice. *Journal of Advertising*, *10*(2), 9-15.

Singh, S.N. & Dalal, N.P. (1999). Web homepages as advertisements. *Communications of the ACM*, *42*(8), 91-98.

Soley, L.C. & Reid, L.N. (1983). Is the perception of informativeness determined by the quantity or the type of information in advertising? *Current Issues and Research in Advertising*, 241-251.

Stern, B.L. & Resnik, A. (1991). Information content in television advertising: a replication and extension. *Journal of Advertising Research*, *31*(2), 36-46.

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, *42*(4), 73-93.

Vijayasarathy, L.R. (2003). Psychographic profiling of the online shopper. *Journal of Electronic Commerce in Organizations*, *1*(3), 48-72.

Westland, J.C. & Au, G. (1998). A comparison of shopping experience across three competing digital retailing interfaces. *International Journal of Electronic Commerce*, *2*(2), 57-69.

Xia, L. & Sudharshan, D. (2000). *An examination of the effects of cognitive interruptions on consumer on-line decision processes.* Paper presented at the Second Marketing Science and the Internet Conference, USC, Los Angeles, April 28-30.

Ylikoski, T. (1994). Cognitive effects of information content in advertising. *Finnish Journal of Business Economics*, *2*, accessed at *www.hkkk.fi/~teylikos/cognitive_effects.htm*

Zeithaml, V.A. (1988). Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. *Journal of Marketing*, *52*, 2-22.

KEY TERMS

Attitude Toward the Ad (Aad): A mediator of advertising response that influences brand attitude and purchase intentions.

Attitude Toward the Site (Ast): A Web user's predisposition to respond either favorably or unfavorably to a website in a natural exposure situation.

Electronic Commerce (EC): The use of computer networks for business communications and commercial transactions

Entertainment: Something that fulfills a visitor's need for aesthetic enjoyment, escapism, diversion, or emotional release.

Informativeness: A Web site's ability to inform consumers of product alternatives for their greatest possible satisfaction.

Interactive Advertising: Advertising that simulates a one-on-one interaction to give consumers more control over their experience with product information than do traditional media ads.

Interactivity: A characteristic of a medium in which the user can influence the form and content of the mediated presentation or experience.

Irritation: An unwanted user feeling caused by tactics perceived to be annoying, offensive, insulting, or overly manipulative.

Site Organization: The structure and navigational ease of a Web site.

Web Marketing: The dissemination of information, promotion of products and services, execution of sales transactions, and enhancement of customer support via a company's Web site.

Content Repurposing for Small Devices

Neil C. Rowe
U.S. Naval Postgraduate School, USA

INTRODUCTION

Content repurposing is the reorganizing of data for presentation on different display hardware (Singh, 2004). It has been particularly important recently with the growth of handheld devices such as personal digital assistants (PDAs), sophisticated telephones, and other small specialized devices. Unfortunately, such devices pose serious problems for multimedia delivery. With their tiny screens (150 by 150 for a basic Palm PDA or 240 by 320 for a more modern one, vs. 640 by 480 for standard computer screens), one cannot display much information (i.e., most of a Web page); with their low bandwidths, one cannot display video and audio transmissions from a server (i.e., streaming) with much quality; and with their small storage capabilities, large media files cannot be stored for later playback. Furthermore, new devices and old ones with new characteristics have been appearing at a high rate, so software vendors are having difficulty keeping pace. So some real-time, systematic, and automated planning could be helpful in figuring how to show desired data, especially multimedia, on a broad range of devices.

BACKGROUND

The World Wide Web is the de facto standard for providing easily accessible information to people. So it is desirable to use it and its language—HTML—as a basis for display for small handheld devices. This would enable people to look up ratings of products while shopping, check routes while driving, and perform knowledge-intensive jobs while walking. HTML is, in fact, device-independent. It requires the display device and its Web-browser software to make decisions about how to display its information within guidelines. But HTML does not provide enough information to devices to ensure much user-friendliness of the resulting display: It does not tell the browser where to break lines or which graphics to keep collocated. Display problems are exacerbated when screen sizes, screen shapes, audio capabilities, or video capabilities are significantly different. Microbrowser markup languages like WML, S-HTML, and HDML, which are based on HTML but designed to better serve the needs of small devices, help, but these only solve some of the problems.

Content repurposing is a general term for reformatting information for different displays. It occurs frequently with content management for an organization's publications (Boiko, 2002), where content or information is broken into pieces and entered in a repository to be used for different publications. However, a repository is not cost-effective unless the information is reused many times, something not generally true for Web pages. Content repurposing for small devices also involves real-time decisions about priorities. For these reasons, the repository approach often is not used with small devices.

Content repurposing can be done either before or after a request for it. Preprocessing can create separate pages for different devices, and the device fetches the page appropriate to it. It also can involve conditional statements in pages that cause different code to be executed for different devices; such statements can be done with code in JavaScript, PHP embedded within HTML, or more complex server codes using such facilities as Java Server Pages (JSP) and Active Server Pages (ASP). It also can involve device-specific planning (Karadkar, 2004). Many popular Web sites provide preprocessed pages for different kinds of devices. Preprocessing is cost-effective for frequently needed content but requires setup time and can require considerable storage space, if there is a large amount of content and ways to display it.

Content repurposing also can be either client-side or server-side. Server-side means a server supplies repurposed information for the client device; client-side means the device itself decides what to display and how. Server-side repurposing saves work for the

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

device, which is important for primitive devices, and can adjust to fluctuations in network bandwidth (Lyu et al., 2003) but requires added complexity in the server and significant time delays in getting information to the server. Devices can have designated proxy servers for their needs. Client-side repurposing, on the other hand, can respond quickly to changing user needs. Its disadvantages are the additional processing burden on an already-slow device and higher bandwidth demands, since information is not eliminated until after it reaches the device. The limitations of small devices require most audio and video repurposing to be server-side.

METHODS OF CONTENT REPURPOSING

Repurposing Strategies

Content repurposing for small devices can be accomplished by several methods, including panning, zooming, reformatting, substitution of links, and modification of content.

A default repurposing method of Internet Explorer and Netscape browser software is to show a window on the full display when it is too large to fit on the device screen. Then the user can manipulate slider bars on the bottom and side of the window to view all the content (pan over it). Some systems break content into overlapping tiles (Kasik, 2004), precomputed units of display information, and users can pan only from tile to tile; this can prevent splitting of key features like buttons and simplifies client-side processing, but it only works for certain kinds of content. Panning may be unsatisfactory for large displays like maps, since considerable screen manipulation may be required, and good understanding may require an overview. But it works fine for most content.

Another idea is to change the scale of view, zooming in (closer) or out (further). This can be either automatic or user-controlled. The MapQuest city-map utility (www.mapquest.com) provides user-controlled zooming by dynamically creating maps at several levels of detail, so the user can start with a city and progressively narrow on a neighborhood (as well as do panning). A problem for zooming out is that some details like text and thin lines cannot be shrunk beyond a certain minimum size and still remain legible. Such details may be optional; for instance, MapQuest omits most street names and many of the streets in its broadest view. But this may not be what the user wants. Different details can be shrunk at different rates, so that lines one pixel wide are not shrunk at all (Ma & Singh, 2003), but this requires content-specific tailoring.

The formatting of the page can be modified to use equivalent constructs that display better on a destination device (Government of Canada, 2004). For instance, with HTML, the fonts can be made smaller or narrower (taking into account viewability on the device) by font tags, line spacing can be reduced, or blank space can be eliminated. Since tables take extra space, they can be converted into text. Small images or video can substitute for large images or video, when their content permits. Text can be presented sequentially in the same box in the screen to save display space (Wobbrock et al., 2002). For audio and video, the sampling or frame rate can be decreased (one image per second is fine for many applications, provided the rate is steady). Visual clues can be added to the display to indicate items just offscreen (Baudisch & Rosenholtz, 2003).

Clickable links can point to blocks of less important information, thereby reducing the amount of content to be displayed at once. This is especially good for media objects, which can require both bandwidth and screen size, but also helps for paragraphs of details. Links can be thumbnail images, which is helpful for pages familiar to the user. Links also can point to pages containing additional links so the scheme can be hierarchical. In fact, Buyukkoten et al. (2002) experimented with repurposing displays containing links exclusively. But insertion of links requires rating the content of the page by importance, a difficult problem in general (as discussed later), to decide what content is converted into links. It also requires a careful wording of text links since just something like "picture here" is not helpful, but a too-long link may be worse than no link at all. Complex link hierarchies also may cause users to get lost.

One also can modify the content of a display by just eliminating unimportant or useless detail and rearranging the display (Gupta et al., 2003). For instance, advertisements, acknowledgements, and horizontal bars can be removed, as well as JavaScript code and Macromedia Flash (SWF) images, since most are only decorative. Removed content need not

be contiguous, as with removal of a power subsystem from a system diagram. In addition, forms and tables can lose their associated graphics. The lines in block diagrams often can be shortened when their lengths do not matter. Color images can be converted to black and white, although one must be careful to maintain feature visibility, perhaps by exaggerating the contrast. User assistance in deciding what to eliminate or summarize is helpful as user judgment provides insights that cannot easily be automated, as with selection of highlights for video (Pea et al., 2004). An important special application is selection of information from a page for each user in a set of users (Han, Perret, & Naghshineh, 2000). Appropriate modification of the display for a mobile device also can be quite radical; for instance, a good way to support route-following on a small device could be to give spoken directions rather than a map (Kray et al., 2003).

Content Rating by Importance

Several of the techniques mentioned above require judgment as to what is important in the data to be displayed. The difficulty of automating this judgment varies considerably with the type of data.

Many editing tools mark document components with additional information like style tags, often in a form compatible with the XML language. This information can assign additional categories to information beyond those of HTML, like identifying text as a introduction, promotion, abstract, author biography, acknowledgements, figure caption, links menu, or reference list (Karben, 1999). These categories can be rated in importance by content-repurposing software, and only text of the top-rated categories shown when display space is tight. Such categorization is especially helpful with media objects (Obrenovic, Starcevic & Selic, 2004), but their automatic content analysis is difficult, and it helps to persuade people to categorize them at least partially.

In the absence of explicit tagging, methods of automatic text summarization from natural language processing can be used. This technology, useful for building digital libraries, can be adapted for the content repurposing problem to display an inferred abstract of a page. One approach is to select sentences from a body of text that are the most important, as measured by various metrics (Alam et al., 2003; McDonald & Chen, 2002), like titles and section headings, first sentences of paragraphs, and distinctive keywords. Keywords alone may suffice to summarize text when the words are sufficiently distinctive (Buyukkotenen et al., 2002). Distinctiveness can be measured by classic measure of TF-IDF, which is $K \log_2 (N/n)$, where K is the number of occurrences of the word in the document or text to be summarized, N is a sample of documents, and n is the number of those documents in that sample having the word at least once. Other useful input for text summarization is the headings of pages linked to (Delort, Bouchon-Meunier, & Rifqi, 2003), since neighbor pages provide content clues. Content also can be classified into semantic units by aggregating clues or even by parsing the page display. For instance, the @ symbol suggests a paragraph of contact information.

Media objects pose more serious problems than text, however, since they can require large bandwidths to download, and images can require considerable display space. In many cases, the media can be inferred to be decorative and can be eliminated (i.e., many banners and sidebars on pages, background sounds). The following simple criteria can distinguish decorative graphics from photographs (Rowe, 2002): size (photographs are larger), frequency of the most common color (graphics have a higher frequency), number of different colors (photographs have more), extremeness of the colors (graphics are more likely to have pure colors), and average variation in color between adjacent pixels in the image (photographs have less). Hu and Bagga (2004) extend this to classify images in order of importance as story, preview, host, commercial, icons and logos, headings, and formatting. Images can be rated by these methods; then, only the top-rated images display until sufficient to fill the screen. Such rating methods are rarely necessary for video and audio, which are almost always accessed by explicit links. Planning can be done on the server for efficient delivery (Chandra, Ellis, & Vahdat, 2000), and the most important media objects can be delivered first.

In some cases, preprocessing can analyze the content of the media object and extract the most representative parts. Video is a good example, because it is characterized by much frame-to-frame redundancy. A variety of techniques can extract representative frames (e.g., one per shot) that convey the gist of the video and reduce the display to a

slide show. If an image is graphics containing subobjects, then the less important subobjects can be removed and a smaller image constructed. An example is a block diagram where text outside the boxes represents notes that can be deleted. Heuristics useful for finding important subobjects are nearby labels, objects at ends of long lines, and adjacent blank areas (Kasik, 2004). In some applications, processing also can do visual abstraction where, for instance, a rectangle is substituted for a complex part of the diagram that is known to be a conceptual unit (Egyed, 2002).

Redrawing the Display

Many of methods discussed require changing the layout of a page of information. Thus, content repurposing needs to use methods of efficient and user-friendly display formatting (Kamada & Kawai, 1991; Tan, Ong, & Wong, 1993). This can be a difficult constraint optimization problem where the primary constraints are those of keeping related information together as much as possible in the display. Examples of what needs to be kept together are section headings with their subsequent paragraphs, links with their describing paragraphs, images with their captions, and images with their text references. Some of the necessary constraints, including device-specific ones, can be learned from observing users (Anderson, Domingos, & Weld, 2001). Even with good page design, content search tools are helpful with large displays like maps to enable users to find things quickly without needing to pan or zoom.

FUTURE WORK

Content repurposing is currently an active area of research, and we are likely to see a number of innovations in the near future in both academia and industry. The large number of competing approaches will dwindle as concensus standards are reached for some of the technology, much as de facto standards have emerged in Web-page style. It is likely that manufacturers of small devices will provide increasingly sophisticated repurposing in their software to reduce the burden on servers. XML increasingly will be used to support repurposing, as it has achieved widespread acceptance in a short time for many other applications. XML will be used to provide standard descriptors for information objects within organizations. But XML will not solve all problems, and the issue of incompatible XML taxonomies could impede progress.

CONCLUSION

Content repurposing recently has become a key issue in management of small wireless devices as people want to display the information they can display on traditional screens and have discovered that it often looks bad on a small device. So strategies are being devised to modify display information for these devices. Simple strategies are effective for some content, but there are many special cases of information that require more sophisticated methods due to their size or organization.

REFERENCES

Alam, H., et al. (2003). Web page summarization for handheld devices: A natural language approach. *Proceedings of 7th International Conference on Document Analysis and Recognition*, Edinburgh, Scotland.

Anderson, C., Domingos, P., & Weld, D. (2001). Personalizing Web sites for mobile users. *Proceedings of 10th International Conference on the World Wide Web*, Hong Kong, China.

Baudisch, P., & Rosenholtz, R. (2003). Halo: A technique for visualizing off-screen objects. *Proceedings of the Conference on Human Factors in Computing Systems*, Ft. Lauderdale, Florida.

Boiko, B. (2002). *Content management bible*. New York: Hungry Minds.

Buyukkokten, O., Kaljuvee, O., Garcia-Molina, H., Paepke, A., & Winograd, T. (2002). Efficient Web browsing on handheld devices using page and form summarization. *ACM Transactions on Information Systems, 20*(1), 82-115.

Chandra, S., Ellis, C., & Vahdat, A., (2000). Application-level differentiated multimedia Web services

using quality aware transcoding. *IEEE Journal on Selected Areas in Communications, 18*(12), 2544-2565.

Delort, J.-Y., Bouchon-Meunier, B., & Rifqi, M. (2003). Enhanced Web document summarization using hyperlinks. *Proceedings of 14th ACM Conference on Hypertext and Hypermedia,* Nottingham, UK.

Egyed, A. (2002). Automatic abstraction of class diagrams. *IEEE Transactions on Software Engineering and Methodology, 11*(4), 449-491.

Government of Canada (2004). Tip sheets: Personal digital assistants (PDA). Retrieved May 5, 2004, from *www.chin.gc.ca/English/Digital_Content/Tip_Sheets/Pda*

Gupta, S., Kaiser, G., Neistadt, D., & Grimm, P. (2003). DOM-based content extraction of HTML documents. *Proceedings of 12th International Conference on the World Wide Web*, Budapest, Hungary.

Han, R., Perret, V., & Naghshineh, M. (2000). WebSplitter: A unified XML framework for multi-device collaborative Web browsing. *Proceedings of ACM Conference on Computer Supported Cooperative Work*, Philadelphia, Pennsylvania.

Hu, J., & Bagga, A. (2004). Categorizing images in Web documents. *IEEE Multimedia, 11*(1), 22-30.

Jing, H., & McKeown, K. (2000). Cut and paste based text summarization. *Proceedings of First Conference of North American Chapter of the Association for Computational Linguistics*, Seattle, Washington.

Kamada, T., & Kawai, S. (1991, January). A general framework for visualizing abstract objects and relations. *ACM Transactions on Graphics, 10*(1), 1-39.

Karadkar, U. (2004). Display-agnostic hypermedia. *Proceedings of 15th ACM Conference on Hypertext and Hypermedia*, Santa Cruz, California.

Karben, A. (1999). News you can reuse—Content repurposing at the *Wall Street Journal* Interactive Edition. *Markup Languages: Theory & Practice, 1*(1), 33-45.

Kasik, D. (2004). Strategies for consistent image partitioning. *IEEE Multimedia, 11*(1), 32-41.

Kray, C., Elting, C., Laakso, K., & Coors, V. (2003). Presenting route instructions on mobile devices. *Proceedings of 8th International Conference on Intelligent User Interfaces*, Miami, Florida.

Lyu, M., Yen, J., Yau, E., & Sze, S. (2003). A wireless handheld multi-modal digital video library client system. *Proceedings of 5th ACM International Workshop on Multimedia Information Retrieval*, Berkeley, California.

Ma, R.-H., & Singh, G. (2003). Effective and efficient infographic image downscaling for mobile devices. *Proceedings of 4th International Workshop on Mobile Computing*, Rostock, Germany.

McDonald, D., & Chen, H. (2002). Using sentence-selection heuristics to rank text in XTRACTOR. *Proceedings of the ACM-IEEE Joint Conference on Digital Libraries*, Portland, Oregon.

Obrenovic, Z., Starcevic, D., & Selic, B. (2004). A model-driven approach to content repurposing. *IEEE Multimedia, 11*(1), 62-71.

Pea, R., Mills, M., Rosen, J., & Dauber, K. (2004). The DIVER project: Interactive digital video repurposing. *IEEE Multimedia, 11*(1), 54-61.

Rowe, N. (2002). MARIE-4: A high-recall, self-improving Web crawler that finds images using captions. *IEEE Intelligent Systems, 17*(4), 8-14.

Singh, G. (2004). Content repurposing. *IEEE Multimedia, 11*(1), 20-21.

Tan, K., Ong, G., & Wong, P. (1993). A heuristics approach to automatic data flow diagram layout. *Proceedings of 6th International Workshop on Computer-Aided Software Engineering*, Singapore.

Wobbrock, J., Forlizzi, J., Hudson, S., & Myers, B. (2002). WebThumb: interaction techniques for small-screen browsers. *Proceedings of 15th ACM Symposium on User Interface Software and Technology*, Paris, France.

KEY TERMS

Content Management: Management of Web pages as assisted by software; Web page bureaucracy.

Content Repurposing: Reorganizing or modifying the content of a graphical display to fit effectively on a different device than its original target.

Key Frames: Representative shots extracted from a video that illustrate its main content.

Microbrowser: A Web browser designed for a small device.

Pan: Move an image window with respect to the portion of the larger image from which it is taken.

PDA: Personal Digital Assistant, a small electronic device that functions like a notepad.

Streaming: Sending multimedia data to a client device at a rate the enables it to be played without having to store it.

Tag: HTML and XML markers that delimit semantically meaningful units in their code.

XML: Extensible Markup Language, a general language for structuring information on the Internet for use with the HTTP protocol, an extension of HTML.

Zoom: Change the fraction of an image being displayed when that image is taken from a larger one.

Content-Based Multimedia Retrieval

Chia-Hung Wei
University of Warwick, UK

Chang-Tsun Li
University of Warwick, UK

INTRODUCTION

In the past decade, there has been rapid growth in the use of digital media such as images, video, and audio. As the use of digital media increases, effective retrieval and management techniques become more important. Such techniques are required to facilitate the effective searching and browsing of large multimedia databases.

Before the emergence of content-based retrieval, media was annotated with text, allowing the media to be accessed by text-based searching (Feng et al., 2003). Through textual description, media can be managed, based on the classification of subject or semantics. This hierarchical structure allows users to easily navigate and browse, and can search using standard Boolean queries. However, with the emergence of massive multimedia databases, the traditional text-based search suffers from the following limitations (Djeraba, 2003; Shah et al., 2004):

- Manual annotations require too much time and are expensive to implement. As the number of media in a databases grows, the difficulty finding desired information increases. It becomes infeasible to manually annotate all attributes of the media content. Annotating a 60-minute video containing more than 100,000 images consumes a vast amount of time and expense.
- Manual annotations fail to deal with the discrepancy of subjective perception. The phrase "a picture is worth a thousand words" implies that the textual description is not sufficient for depicting subjective perception. Capturing all concepts, thoughts, and feelings for the content of any media is almost impossible.
- Some media contents are difficult to describe concretely in words. For example, a piece of melody without lyrics or an irregular organic shape cannot be expressed easily in textual form, but people expect to search media with similar contents based on examples they provide. In an attempt to overcome these difficulties, content-based retrieval employs content information to automatically index data with minimal human intervention.

APPLICATIONS

Content-based retrieval has been proposed by different communities for various applications. These include:

- **Medical Diagnosis:** The amount of digital medical images used in hospitals has increased tremendously. As images with the similar pathology-bearing regions can be found and interpreted, those images can be applied to aid diagnosis for image-based reasoning. For example, Wei & Li (2004) proposed a general framework for content-based medical image retrieval and constructed a retrieval system for locating digital mammograms with similar pathological parts.
- **Intellectual Property:** Trademark image registration has applied content-based retrieval techniques to compare a new candidate mark with existing marks to ensure that there is no repetition. Copyright protection also can benefit from content-based retrieval, as copyright owners are able to search and identify unauthorized copies of images on the Internet. For example, Wang & Chen (2002) developed a content-based system using hit statistics to retrieve trademarks.
- **Broadcasting Archives:** Every day, broadcasting companies produce a lot of audiovisual data. To deal with these large archives, which can

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

contain millions of hours of video and audio data, content-based retrieval techniques are used to annotate their contents and summarize the audiovisual data to drastically reduce the volume of raw footage. For example, Yang et al. (2003) developed a content-based video retrieval system to support personalized news retrieval.

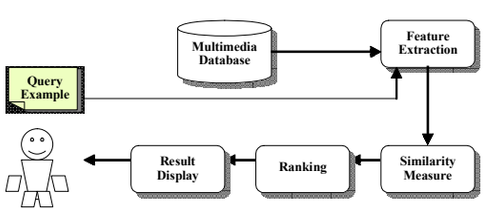
- **Information Searching on the Internet:** A large amount of media has been made available for retrieval on the Internet. Existing search engines mainly perform text-based retrieval. To access the various media on the Internet, content-based search engines can assist users in searching the information with the most similar contents based on queries. For example, Hong & Nah (2004) designed an XML scheme to enable content-based image retrieval on the Internet.

DESIGN OF CONTENT-BASED RETRIEVAL SYSTEMS

Before discussing design issues, a conceptual architecture for content-based retrieval is introduced and illustrated in Figure 1.

Content-based retrieval uses the contents of multimedia to represent and index the data (Wei & Li, 2004). In typical content-based retrieval systems, the contents of the media in the database are extracted and described by multi-dimensional feature vectors, also called descriptors. The feature vectors of the media constitute a feature dataset. To retrieve desired data, users submit query examples to the retrieval system. The system then represents these examples with feature vectors. The distances (i.e., similarities) between the feature vectors of the query example and those of the media in the feature dataset are then computed and ranked. Retrieval is conducted by applying an indexing scheme to provide an efficient way to search the media database. Finally, the system ranks the search results and then returns the top search results that are the most similar to the query examples.

Figure 1. A conceptual architecture for content-based retrieval



For the design of content-based retrieval systems, a designer needs to consider four aspects: feature extraction and representation, dimension reduction of feature, indexing, and query specifications, which will be introduced in the following sections.

FEATURE EXTRACTION AND REPRESENTATION

Representation of media needs to consider which features are most useful for representing the contents of media and which approaches can effectively code the attributes of the media. The features are typically extracted off-line so that efficient computation is not a significant issue, but large collections still need a longer time to compute the features. Features of media content can be classified into low-level and high-level features.

Low-Level Features

Low-level features such as object motion, color, shape, texture, loudness, power spectrum, bandwidth, and pitch are extracted directly from media in the database (Djeraba, 2002). Features at this level are objectively derived from the media rather than referring to any external semantics. Features extracted at this level can answer queries such as "finding images with more than 20% distribution in blue and green color," which might retrieve several images with blue sky and green grass (see Picture 1). Many effective approaches to low-level feature extraction have been developed for various purposes (Feng et al., 2003; Guan et al., 2001).

High-Level Features

High-level features are also called semantic features. Features such as timbre, rhythm, instruments, and events involve different degrees of semantics contained in the media. High-level features are supposed

to deal with semantic queries (e.g.,"finding a picture of water" or "searching for Mona Lisa Smile"). The latter query contains higher-degree semantics than the former. As water in images displays the homogeneous texture represented in low-level features, such a query is easier to process. To retrieve the latter query, the retrieval system requires prior knowledge that can identify that Mona Lisa is a woman, who is a specific character rather than any other woman in a painting.

The difficulty in processing high-level queries arises from external knowledge with the description of low-level features, known as the semantic gap. The retrieval process requires a translation mechanism that can convert the query of "Mona Lisa Smile" into low-level features. Two possible solutions have been proposed to minimize the semantic gap (Marques & Furht, 2002). The first is automatic metadata generation to the media. Automatic annotation still involves the semantic concept and requires different schemes for various media (Jeon et al., 2003). The second uses relevance feedback to allow the retrieval system to learn and understand the semantic context of a query operation. Relevance feedback will be discussed in the Relevance Feedback section.

DIMENSION REDUCTION OF FEATURE VECTOR

Many multimedia databases contain large numbers of features that are used to analyze and query the database. Such a feature-vector set is considered as high dimensionality. For example, Tieu & Viola (2004) used over 10,000 features of images, each describing a local pattern. High dimensionality causes the "curse of dimension" problem, where the complexity and computational cost of the query increases exponentially with the number of dimensions (Egecioglu et al., 2004). Dimension reduction is a popular technique to overcome this problem and support efficient retrieval in large-scale databases. However, there is a tradeoff between the efficiency obtained through dimension reduction and the completeness obtained through the information extracted. If each data is represented by a smaller number of dimensions, the speed of retrieval is increased. However, some information may be lost. One of the most widely used techniques in multimedia retrieval is Principal Component Analysis (PCA). PCA is used to transform the original data of high dimensionality into a new coordinate system with low dimensionality by finding data with high discriminating power. The new coordinate system removes the redundant data and the new set of data may better represent the essential information. Shyu et al. (2003) presented an image database retrieval framework and applied PCA to reduce the image feature vectors.

Picture 1. There are more than 20% distributions in blue and green color in this picture

INDEXING

The retrieval system typically contains two mechanisms: similarity measurement and multi-dimensional indexing. Similarity measurement is used to find the most similar objects. Multi-dimensional indexing is used to accelerate the query performance in the search process.

Similarity Measurement

To measure the similarity, the general approach is to represent the data features as multi-dimensional points and then to calculate the distances between the corresponding multi-dimensional points (Feng et al., 2003). Selection of metrics has a direct impact on the performance of a retrieval system. Euclidean distance is the most common metric used to measure the distance between two points in multi-dimensional space (Qian et al., 2004). However, for some applications, Euclidean distance is not compatible with the human perceived similarity. A number of metrics (e.g., Mahalanobis Distance, Minkowski-Form Dis-

tance, Earth Mover's Distance, and Proportional Transportation Distance) have been proposed for specific purposes. Typke et al. (2003) investigated several similarity metrics and found that Proportional Transportation Distance fairly reflected melodic similarity.

Multi-Dimensional Indexing

Retrieval of the media is usually based not only on the value of certain attributes, but also on the location of a feature vector in the feature space (Fonseca & Jorge, 2003). In addition, a retrieval query on a database of multimedia with multi-dimensional feature vectors usually requires fast execution of search operations. To support such search operations, an appropriate multi-dimensional access method has to be used for indexing the reduced but still high dimensional feature vectors. Popular multi-dimensional indexing methods include R-tree (Guttman, 1984) and R*-tree (Beckmann et al., 1990). These multi-dimensional indexing methods perform well with a limit of up to 20 dimensions. Lo & Chen (2002) proposed an approach to transform music into numeric forms and developed an index structure based on R-tree for effective retrieval.

QUERY SPECIFICATIONS

Querying is used to search for a set of results with similar content to the specified examples. Based on the type of media, queries in content-based retrieval systems can be designed for several modes (e.g., query by sketch, query by painting [for video and image], query by singing [for audio], and query by example). In the querying process, users may be required to interact with the system in order to provide relevance feedback, a technique that allows users to grade the search results in terms of their relevance. This section will describe the typical query by example mode and discuss the relevance feedback.

Query by Example

Queries in multimedia retrieval systems are traditionally performed by using an example or series of examples. The task of the system is to determine which candidates are the most similar to the given example. This design is generally termed Query By Example (QBE) mode. The interaction starts with an initial selection of candidates. The initial selection can be randomly selected candidates or meaningful representatives selected according to specific rules. Subsequently, the user can select one of the candidates as an example, and the system will return those results that are most similar to the example. However, the success of the query in this approach heavily depends on the initial set of candidates. A problem exists in how to formulate the initial panel of candidates that contains at least one relevant candidate. This limitation has been defined as page zero problem (La Cascia et al., 1998). To overcome this problem, various solutions have been proposed for specific applications. For example, Sivic and Zisserman (2004) proposed a method that measures the reoccurrence of spatial configurations of viewpoint invariant features to obtain the principal objects, characters, and scenes, which can be used as entry points for visual search.

Relevance Feedback

Relevance feedback was originally developed for improving the effectiveness of information retrieval systems. The main idea of relevance feedback is for the system to understand the user's information needs. For a given query, the retrieval system returns initial results based on predefined similarity metrics. Then, the user is required to identify the positive examples by labeling those that are relevant to the query. The system subsequently analyzes the user's feedback using a learning algorithm and returns refined results. Two of the learning algorithms frequently used to iteratively update the weight estimation were developed by Rocchio (1971) and Rui and Huang (2002).

Although relevance feedback can contribute retrieval information to the system, two challenges still exist: (1) the number of labeled elements obtained through relevance feedback is small when compared to the number of unlabeled in the database; (2) relevance feedback iteratively updates the weight of high-level semantics but does not automatically modify the weight for the low-level features. To solve these problems, Tian et al. (2000) proposed an approach for combining unlabeled data in supervised learning to achieve better classification.

FUTURE RESEARCH ISSUES AND TRENDS

Since the 1990s, remarkable progress has been made in theoretical research and system development. However, there are still many challenging research problems. This section identifies and addresses some issues in the future research agenda.

Automatic Metadata Generation

Metadata (data about data) is the data associated with an information object for the purposes of description, administration, technical functionality, and so on. Metadata standards have been proposed to support the annotation of multimedia content. Automatic generation of annotations for multimedia involves high-level semantic representation and machine learning to ensure accuracy of annotation. Content-based retrieval techniques can be employed to generate the metadata, which can be used further by the text-based retrieval.

Establishment of Standard Evaluation Paradigm and Test-Bed

The National Institute of Standards and Technology (NIST) has developed TREC (Text REtrieval Conference) as the standard test-bed and evaluation paradigm for the information retrieval community. In response to the research needs from the video retrieval community, the TREC released a video track in 2003, which became an independent evaluation (called TRECVID) (Smeaton, 2003). In music information retrieval, a formal resolution expressing a similar need was passed in 2001, requesting a TREC-like standard test-bed and evaluation paradigm (Downie, 2003). The image retrieval community still awaits the construction and implementation of a scientifically valid evaluation framework and standard test bed.

Embedding Relevance Feedback

Multimedia contains large quantities of rich information and involves the subjectivity of human perception. The design of content-based retrieval systems has turned out to emphasize an interactive approach instead of a computer-centric approach. A user interaction approach requires human and computer to interact in refining the high-level queries. Relevance feedback is a powerful technique used for facilitating interaction between the user and the system. The research issue includes the design of the interface with regard to usability and learning algorithms, which can dynamically update the weights embedded in the query object to model the high-level concepts and perceptual subjectivity.

Bridging the Semantic Gap

One of the main challenges in multimedia retrieval is bridging the gap between low-level representations and high-level semantics (Lew & Eakins, 2002). The semantic gap exists because low-level features are more easily computed in the system design process, but high-level queries are used at the starting point of the retrieval process. The semantic gap is not only the conversion between low-level features and high-level semantics, but it is also the understanding of contextual meaning of the query involving human knowledge and emotion. Current research intends to develop mechanisms or models that directly associate the high-level semantic objects and representation of low-level features.

CONCLUSION

The main contributions in this article were to provide a conceptual architecture for content-based multimedia retrieval, to discuss the system design issues, and to point out some potential problems in individual components. Finally, some research issues and future trends were identified and addressed.

The ideal content-based retrieval system from a user's perspective involves the semantic level. Current content-based retrieval systems generally make use of low-level features. The semantic gap has been a major obstacle for content-based retrieval. Relevance feedback is a promising technique to bridge this gap. Due to the efforts of the research community, a few systems have started to employ high-level features and are able to deal with some semantic queries. Therefore, more intelligent content-based retrieval systems can be expected in the near future.

REFERENCES

Beckmann, N., Kriegel, H.-P., Schneider, R., & Seeger, B. (1990). The R*-tree: An efficient and robust access method for points and rectangles. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Atlantic City, NJ, USA.

Djeraba, C. (2002). Content-based multimedia indexing and retrieval. *IEEE MultiMedia, 9*(2) 18-22.

Djeraba, C. (2003). Association and content-based retrieval. *IEEE Transactions on Knowledge and Data Engineering, 15*(1), 118-135.

Downie, J.S. (2003). Toward the scientific evaluation of music information retrieval systems. *Proceedings of the Fourth International Symposium on Music Information Retrieval*, Washington, D.C., USA.

Egecioglu, O., Ferhatosmanoglu, H., & Ogras, U. (2004). Dimensionality reduction and similarity computation by inner-product approximations. *IEEE Transactions on Knowledge and Data Engineering,16*(6), 714-726.

Feng, D., Siu, W.C., & Zhang, H.J. (Eds.). (2003). *Multimedia information retrieval and management:Technological fundamentals and applications*. Berlin: Springer.

Fonseca, M.J., & Jorge, J.A. (2003). Indexing highdimensional data for content-based retrieval in large database. *Proceedings of the Eighth International Conference on Database Systems for Advanced Applications*, Kyoto, Japan.

Guan, L., Kung S.-Y., & Larsen, J. (Eds.). (2001). *Multimedia image and video processing*. New York: CRC Press.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Boston, MA, USA.

Hong, S., & Nah, Y. (2004). An intelligent image retrieval system using XML. *Proceedings of the 10th International Multimedia Modelling Conference,* Brisbane, Australia.

Jeon, J., Lavrenko, V., & Manmatha, R. (2003). Automatic image annotation and retrieval using crossmedia relevance models. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada.

La Cascia, M., Sethi, S., & Sclaroff, S. (1998). Combining textual and visual cues for content-based image retrieval on the World Wide Web. *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa Barbara, CA, USA.

Lew, M.S., Sebe, N., & Eakins, J.P. (2002). Challenges of image and video retrieval. *Proceedings of the International Conference on Image and Video Retrieval, Lecture Notes in Computer Science*, London, UK.

Lo, Y.-L., & Chen, S.-J. (2002). The numeric indexing for music data. *Proceedings of the 22nd International Conference on Distributed Computing Systems Workshops*. Vienna, Austria.

Marques, O., & Furht, B. (2002). *Content-based image and video retrieval.* London: Kluwer.

Qian, G., Sural, S., Gu, Y., & Pramanik, S. (2004). Similarity between Euclidean and cosine angle distance for nearest neighbor queries. *Proceedings of 2004 ACM Symposium on Applied Computing*, Nicosia, Cyprus.

Rocchio, J.J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system—Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice Hall.

Rui, Y., & Huang, T. (2002). Learning based relevance feedback in image retrieval. In A.C. Bovik, C.W. Chen, & D. Goldfof (Eds.), *Advances in image processing and understanding: A festschrift for Thomas S. Huang* (pp. 163-182). New York: World Scientific Publishing.

Shah, B., Raghavan, V., & Dhatric, P. (2004). Efficient and effective content-based image retrieval using space transformation. *Proceedings of the 10th International Multimedia Modelling Conference*, Brisbane, Australia.

Shyu, C.R., et al. (1999). ASSERT: A physician-in-the-loop content based retrieval system for HRCT image databases. *Computer Vision and Image Understanding, 75*(1-2), 111-132.

Sivic, J., & Zisserman, A. (2004). Video data mining using configurations of viewpoint invariant regions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Washington, DC, USA.

Smeaton, A.F., Over, P. (2003). TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. *Proceedings of the International Conference on Image and Video Retrieval*, Urbana, IL, USA.

Tian, Q., Wu, Y., & Huang, T.S. (2000). Incorporate discriminant analysis with EM algorithm in image retrieval. *Proceedings of the IEEE International Conference on Multimedia and Expo,* New York, USA.

Tieu, K., & Viola, P. (2004). Boosting image retrieval. *International Journal of Computer Vision, 56*(1-2), 17-36.

Typke, R., Giannopoulos, P., Veltkamp, R.C. Wiering, F., & Oostrum, R.V. (2003). Using transportation distances for measuring melodic similarity. *Proceedings of the Fourth International Symposium on Music Information Retrieval,* Washington, DC, USA.

Wang, C.-C., & Chen, L.-H. (2002). Content-based color trademark retrieval system using hit statistic. *International Journal of Pattern and Artificial Intelligence,16*(5), 603-619.

Wei, C.-H., & Li, C.-T. (2004). A general framework for content-based medical image retrieval with its application to mammogram retrieval. *Proceedings of IS&T/SPIE International Symposium on Medical Imaging*, San Diego, CA, USA.

Yang, H., Chaisorn, L., Zhao, Y., Neo, S.-Y., & Chua, T.-S. (2003). VideoQA: Question answering on news video. *Proceedings of the Eleventh ACM International Conference on Multimedia,* Berkeley, CA, USA.

KEY TERMS

Boolean Query: A query that uses Boolean operators (AND, OR, and NOT) to formulate a complex condition. A Boolean query example can be "university" OR "college."

Content-Based Retrieval: An application that directly makes use of the contents of media rather than annotation inputted by the human to locate desired data in large databases.

Feature Extraction: A subject of multimedia processing that involves applying algorithms to calculate and extract some attributes for describing the media.

Query by Example: A method of searching a database using example media as search criteria. This mode allows the users to select predefined examples requiring the users to learn the use of query languages.

Relevance Feedback: A technique that requires users to identify positive results by labeling those that are relevant to the query and subsequently analyzes the user's feedback using a learning algorithm.

Semantic Gap: The difference between the high-level user perception of the data and the lower-level representation of the data used by computers. As high-level user perception involves semantics that cannot be translated directly into logic context, bridging the semantic gap is considered a challenging research problem.

Similarity Measure: A measure that compares the similarity of any two objects represented in the multi-dimensional space. The general approach is to represent the data features as multi-dimensional points and then to calculate the distances between the corresponding multi-dimensional points.

Context-Awareness in Mobile Commerce

Jun Sun
Texas A&M University, USA

Marshall Scott Poole
Texas A&M University, USA

INTRODUCTION

Advances in wireless network and multimedia technologies enable mobile commerce (m-commerce) information service providers to know the location and surroundings of mobile consumers through GPS-enabled and camera-embedded cell phones. Context awareness has great potential for creating new service modes and improving service quality in m-commerce. To develop and implement successful context-aware applications in m-commerce, it is critical to understand the concept of the "context" of mobile consumers and how to access and utilize contextual information in an appropriate way. This article dissects the context construct along both the behavioral and physical dimensions from the perspective of mobile consumers, developing a classification scheme for various types of consumer contexts. Based on this classification scheme, it discusses three types of context-aware applications—non-interactive mode, interactive mode and community mode—and describes newly proposed applications as examples of each.

UTILIZING CONSUMER CONTEXT: OPPORTUNITY AND CHALLENGE

M-commerce gets its name from consumers' usage of wireless handheld devices, such as cell phones or PDAs, rather than PCs as in traditional e-commerce (Mennecke & Strader, 2003). Unlike e-commerce users, m-commerce users enjoy a pervasive and ubiquitous computing environment (Lyttinen & Yoo, 2002), and therefore can be called "mobile consumers."

A new generation of wireless handheld devices is embedded or can be connected with GPS receivers, digital cameras and other wearable sensors. Through wireless networks, mobile consumers can share information about their location, surroundings and physiological conditions with m-commerce service providers. Such information is useful in context-aware computing, which employs the collection and utilization of user context information to provide appropriate services to users (Dey, 2001; Moran & Dourish, 2001). The new multimedia framework standard, MPEG-21, describes how to adapt such digital items as user and environmental characteristics for universal multimedia access (MPEG Requirements Group, 2002). Wireless technology and multimedia standards give m-commerce great potential for creating new context-aware applications in m-commerce.

However, user context is a dynamic construct, and any given context has different meanings for different users (Greenberg, 2001). In m-commerce as well, consumer context takes on unique characteristics, due to the involvement of mobile consumers. To design and implement context-aware applications in m-commerce, it is critical to understand the nature of consumer context and the appropriate means of accessing and utilizing different types of contextual information. Also, such an understanding is essential for the identification and adaptation of context-related multimedia digital items in m-commerce.

CONSUMER CONTEXT AND ITS CLASSIFICATION

Dey, Abowd and Salber (2001) defined "context" in context-aware computing as "any information that can be used to characterize the situation of entities (i.e., whether a person, place or object) that are considered relevant to the interaction between a user and an application …" (p. 106). This definition makes it clear that context can be "any information," but it limits context to those things relevant to the behavior of users in interacting with applications.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Most well-known context-relevant theories, such as Situated Action Theory (Suchman, 1987) and Activity Theory (Nardi, 1997), agree that "user context" is a concept inseparable from the goals or motivations implicit in user behavior. For specific users, interacting with applications is the means to their goals rather than an end in itself. User context, therefore, should be defined based on typical user behavior that is identifiable with its motivation.

According to the Merriam-Webster Collegiate Dictionary, the basic meaning of context is "a setting in which something exists or occurs." Because the typical behavior of mobile consumers is consumer behavior, the user context in m-commerce, which we will term *consumer context*, is a setting in which various types of consumer behavior occur.

Need Context and Supply Context

Generally speaking, consumer behavior refers to how consumers acquire and consume goods and services (both informational and non-informational) to satisfy their needs (e.g., Soloman, 2002). Therefore, consumer behavior is, to a large extent, shaped by two basic factors: consumer needs and what is available to meet such needs. Correspondingly, consumer context can be classified conceptually into "need context" and "supply context." A *need context* is composed of stimuli that can potentially arouse a consumer's needs. A *supply context* is composed of resources that can potentially meet a consumer's needs.

This behavioral classification of consumer context is based on perceptions rather than actual physical states, because the same physical context can have different meanings for different consumers. Moreover, a contextual element can be in a consumer's need and supply contexts simultaneously. For example, the smell or sight of a restaurant may arouse a consumer's need for a meal, while the restaurant is part of the supply context. However, it is improper to infer what a consumer needs based on his or her supply context (see below). Therefore, this conceptual differentiation of consumer contexts is important for the implementation of context-aware applications in m-commerce, which should be either need context-oriented or supply context-oriented.

The needs of a consumer at any moment are essential for determining how a context is relevant to the consumer. However, "consumer need" is both a multi-level construct and a personal issue. According to Maslow (1954), human need is a psychological construct composed of five levels: physiological, safety, social, ego and self-actualization. While it is feasible to infer some of the more basic needs of mobile consumers, including physiological and safety needs, based on relevant context information, it is almost impossible to infer other higher-level needs. Moreover, consumer need is a personal issue involving privacy concerns. Because context-aware computing should not violate the personal privacy of users by depriving them of control over their needs and priorities (Ackerman, Darrell & Weitzner, 2001), it is improper to infer a consumer's needs solely based on his or her supply context and provide services accordingly. It is for this reason that pushing supply context information to mobile consumers based on where they are is generally unacceptable to users.

When consumers experience emergency conditions, including medical emergencies and disastrous events, they typically need help from others. Necessary services are usually acceptable to consumers when their urgent "physiological" and "safety" needs can be correctly inferred based on relevant context information. Context-aware applications can stand alert for such need contexts of consumers and provide necessary services as soon as possible when any emergencies occur. Such context-awareness in m-commerce can be denoted as *need-context-awareness*.

Under normal conditions, context-aware applications should let consumers determine their own needs and how certain supply contexts are relevant. The elements of supply contexts, including various sites, facilities and events, usually locate or occur in certain functionally defined areas, such as shopping plazas, tourist parks, traffic systems, sports fields and so on. Information about such contextual elements in certain areas can be gathered from suppliers and/or consumers and stored in databases. *Supply-context-awareness*, therefore, concerns how to select, organize and deliver such information to mobile consumers based on their locations and needs.

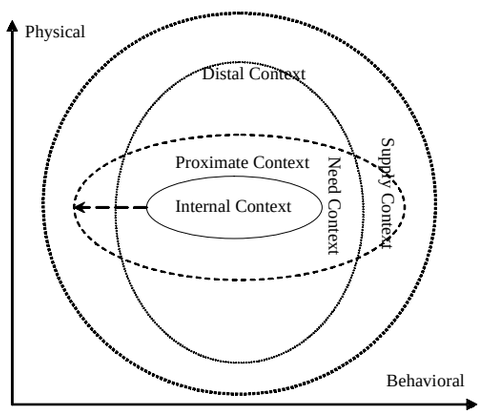
Internal Context, Proximate Context and Distal Context

Besides the behavioral classification, contextual elements can also be classified based on their physical

locus. According to whether the contextual elements are within or outside the body of a consumer, a consumer context can be divided into internal and external contexts. An *internal context* is comprised of sensible body conditions that may influence a consumer's needs. By definition, internal context is part of need context. An *external context*, however, can refer to both the supply context and part of the need context that is outside of a consumer.

According to whether the contextual elements can be directly perceived by a consumer, his or her external context can be divided into "proximate context" and "distal context." A *proximate context* is that part of external context close enough to be directly perceivable to a consumer. A *distal context* is that part of external context outside the direct perception of a consumer. Mobile consumers do not need to be informed of their proximate context, but may be interested in information about their distal context. Context-aware information systems, which are able to retrieve the location-specific context information, can be a source of distal context information for mobile consumers. Besides, consumers can describe or even record information about their proximate context and share it with others through wireless network. To those who are not near the same locations, the information pertains to their distal contexts.

Figure 1. A classification of consumer context



Note: ←– – indicates the range of direct perceivability.

Figure 1 illustrates a classification scheme that combines two dimensions of consumer context, physical and behavioral. The need context covers all the internal context and part of the external context. A subset of need context that can be utilized by need context-aware applications is *emergency context*; includes *internal emergency context,* which comprises urgent physiological conditions (e.g., abnormal heart rate, blood pressure and body temperature); and *external emergency context,* which emerges at the occurrence of natural and human disasters (e.g., tornado, fire and terrorist attacks). The supply context, however, is relatively more stable or predictable, and always external to a consumer. Supply context-aware applications mainly help mobile consumers obtain and share desirable supply context information. This classification scheme provides a guideline for the identification and adaptation of context-related multimedia digital items in m-commerce.

CONTEXT-AWARE APPLICATIONS IN M-COMMERCE

Context-aware applications in m-commerce are applications that obtain, utilize and/or exchange context information to provide informational and/or non-informational services to mobile consumers. They can be designed and implemented in various ways according to their orientation towards either need or supply context, and ways of collecting, handling and delivering context information.

It is generally agreed that location information of users is essential for context-aware computing (e.g., Grudin, 2001). Similarly, context-aware applications in m-commerce need the location information of mobile consumers to determine their external contexts and provide location-related services. Today's GPS receivers can be made very small, and they can be plugged or embedded into wireless handheld devices. Therefore, it is technically feasible for context-aware applications to acquire the location information of mobile consumers. However, it is not ethically appropriate to keep track of the location of consumers all of the time because of privacy concerns. Rather, consumers should be able to determine whether and/or when to release their location information except in emergency conditions.

There can be transmission of contextual information in either direction over wireless networks between the handheld devices of mobile consumers and information systems that host context-aware applications. For applications oriented towards the internal need context, there is at least the flow of physiological and location information from the consumer to the systems. Other context-aware applications typically intend to help mobile consumers get information about their distal contexts and usually involve information flow in both directions.

In this sense, mobile consumers who use context-aware applications are communicating with either information systems or other persons (usually users) through the mediation of systems. For user-system communications, it is commonly believed that the interactivity of applications is largely about whether they empower users to exert control on the content of information they can get from the systems (e.g., Jensen, 1998). Therefore, the communications between a consumer and a context-aware system can be either non-interactive or interactive, depending on whether the consumer can actively specify and choose what context-related information they want to obtain. Accordingly, there are two modes of context-aware applications that involve communication between mobile consumers and information systems: the non-interactive mode and the interactive mode. For user-user communications, context-aware applications mediate the exchange of contextual information among mobile consumers. This represents a third mode: the community mode. This classification of context-aware applications into non-interactive, interactive and community modes is consistent with Bellotti and Edwards' (2001) classification of context awareness into responsiveness to environment, responsiveness to people and responsiveness to the interpersonal. Below, we will discuss these modes and give an example application for each.

Non-Interactive Mode

Successful context-aware applications in m-commerce must cater to the actual needs of mobile consumers. The non-interactive mode of context-aware applications in m-commerce is oriented toward the need context of consumers: It makes assumptions about the needs that mobile consumers have in certain contexts and provides services accordingly. As mentioned above, the only contexts in which it is appropriate to assess consumer needs are certain emergency conditions. We can call non-interactive context-aware applications that provide necessary services in response to emergency contexts Wireless Emergency Services (WES). Corresponding to the internal and external emergency contexts of mobile consumers, there are two types of WES: Personal WES and Public WES.

Personal WES are applications that provide emergency services (usually medical) in response to the internal emergency contexts of mobile consumers. Such applications use bodily attached sensors (e.g., wristwatch-like sensors) to keep track of certain physiological conditions of service subscribers. Whenever a sensor detects anything abnormal, such as a seriously irregular heart rate, it will trigger the wearer's GPS-embedded cell phone to send both location information and relevant physiological information to a relevant emergency service. The emergency service will then send an ambulance to the location and medical personnel can prepare to administer first-aid procedure based on the physiological information and medical history of the patient. The connection between the sensor and cell phone can be established through some short-distance wireless data-communication technology, such as Bluetooth.

Public WES are applications that provide necessary services (mainly informational services) to mobile consumers in response to their external emergency contexts. Such applications stand on alert for any disastrous events in the coverage areas and detect external context information through various fixed or remote sensors or reports by people in affected areas. When a disaster occurs (e.g., tornado), the Public WES systems gather the location information from the GPS-embedded cell phones of those nearby through the local transceivers. Based on user location and disaster information, the systems then give alarms to those involved (e.g., "There are tornado activities within one mile!") and display detailed self-help information, such as evacuation routes and nearby shelters, on their cell phones.

Interactive Mode

The interactive mode of context-aware applications in m-commerce does not infer consumer needs based on contextual information, but lets consumers express

their particular information requirements regarding what they need. Therefore, the interactive mode is not oriented towards the need contexts of consumers, but their supply contexts. The Information Requirement Elicitation (IRE) proposed by Sun (2003) is such an interactive context-aware application.

In the IRE approach, mobile consumers can express their needs by clicking the links on their wireless handheld devices, such as "restaurants" and "directions," that they have pre-selected from a services inventory. Based on such requests, IRE-enabled systems obtain the relevant supply context information of the consumers, and elicit their information requirements with adaptive choice prompts (e.g., food types and transportation modes available). A choice prompt is generated based on the need expressed by a consumer, the supply context and the choice the consumer has made for the previous prompt. When the information requirements of mobile consumers are elicited to the level of specific suppliers they prefer, IRE-enabled systems give detailed supplier information, such as directions and order forms.

The IRE approach allows the consumers to specify which part of their distal supply context they want to know in detail through their interactions with information systems. It attempts to solve the problem of inconvenience in information search for mobile consumers, a key bottle neck in m-commerce. However, it requires consumers to have a clear notion of what they want.

Community Mode

The community mode of context-aware applications in m-commerce mediates contextual information exchange among a group of mobile consumers. Consumers can only share information about what is directly perceivable to them, their proximate contexts. However, the information shared about the proximate context may be interesting distal context information for others if it is relevant to their consumption needs or other interests. A group of mobile consumers in a functionally defined business area have a common supply context, and they may learn about it through sharing context information with each other. Some applications in DoCoMo in Japan have the potential to operate in the community mode.

Wireless Local Community (WLC) is an approach to facilitate the exchange of context information for a group of mobile consumers in a common supply context, such as a shopping plaza, tourist park or sports field (Sun & Poole, working paper). In such an area, mobile consumers with certain needs or interests can join a WLC to share information about their proximate supply contexts with each other (e.g., seeing a bear in a national park). Because the information shared by different consumers is about different parts of the bigger common supply context, the complementary contributions are likely to achieve an "informational synergy." Compared with the IRE approach, the WLC approach allows mobile consumers to obtain potentially useful or interesting context information without indicating what they want.

Table 1 illustrates the primary context orientations of three modes of context-aware applications. The need context-aware applications are usually non-interactive. Personal WES applications are oriented towards the internal need context of mobile consumers, while Public WES applications are oriented towards the external (especially distal) need context

Table 1. Primary context orientations of context-aware applications

Physical / Behavioral	Internal Context	Proximate Context	Distal Context
Need Context	(Personal WES)	←Non-Interactive→	(Public WES)
Supply Context	N/A	Community (WLC)	Interactive (IRE)

of mobile consumers. The supply context-aware applications should be either of the interactive mode or community mode. As an example of interactive mode applications, IRE systems help mobile consumers know the part of their distal supply context they are interested in through choice prompts. As an example of community mode applications, WLC enables mobile consumers to share their proximate supply context with each others.

CONCLUSION

The advance in multimedia standards and network technology endows m-commerce great potential in providing mobile consumers context-aware applications. An understanding of consumer context is necessary for the development of various context-aware applications, as well as the identification and adaptation of context-related multimedia digital items. This article defines dimensions of consumer context and differentiates three modes of context-aware applications in m-commerce: the non-interactive, interactive and community modes. While applications for the interactive and community modes are in rather short supply at present, all indications are that they will burgeon as m-commerce continues to develop. Example applications are given to stimulate the thoughts on developing new applications.

Further technical and behavioral issues must be addressed before the design, implementation and operation of context-aware applications in m-commerce. Such issues may include: network bandwidth and connection, digital elements compatibility, content presentation, privacy protection, interface design, service sustainability and so on. We hope that this article can enhance further discussions in this area.

REFERENCES

Ackerman, M., Darrell, T., & Weitzner, D.J. (2001). Privacy in context. *Human-Computer Interaction, 16*, 167-176.

Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: Human considerations in context-aware systems. *Human-Computer Interaction, 16*, 193-212.

Dey, A.K. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal,* (1), 4-7.

Dey, A.K., Abowd, G.D., & Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction, 16*, 97-166.

Greenberg, S. (2001). Context as a dynamic construct. *Human-Computer Interaction*, 16, 257-268.

Grudin, J. (2001). Desituating action: Digital representation of context. *Human-Computer Interaction, 16,* 269-286.

Jensen, J.F. (1998). Interactivity: tracing a new concept in media and communication studies. *Nordicom Review,* (1), 185-204.

Lyttinen, K., & Yoo, Y. (2002). Issues and challenges in ubiquitous computing. *Communication of the ACM,* (12), 63-65.

Maslow, A.H. (1954). *Motivation and personality.* New York: Harper & Row.

Mennecke, B.E., & Strader, T.J. (2002). *Mobile commerce: Technology, theory and applications.* Hershey, PA: Idea Group Publishing.

Moran, T.P., & Dourish, P. (2001). Introduction to this special issue on context-aware computing. *Human-Computer Interaction, 16*, 87-95.

MPEG Requirements Group. (2002). *MPEG-21 Overview*. ISO/MPEG N5231.

Nardi, B. (1997). *Context and consciousness: Activity theory and human computer interaction.* Cambridge. MA: MIT Press.

Solomon, M.R. (2002). *Consumer behaviour: buying, having, and being* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication.* Cambridge: University Press.

Sun, J. (2003). Information requirement elicitation in mobile commerce. *Communications of the ACM, 46*(12), 45-47.

Sun, J. & Poole, M.S. (working paper). Wireless local community in mobile commerce. Information & Operations Management, Texas A&M University.

KEY TERMS

Consumer Context: The setting in which certain consumer behaviour occurs. It can be can be classified conceptually into "need context" and "supply context," and physically into "internal context," "proximate context" and "distal context."

Distal Context: The physical scope of a consumer context that is outside the direct perception of the consumer. Most context-aware applications intend to help mobile consumers obtain useful and interesting information about their distal context.

Information Requirement Elicitation (IRE): An interactive mode of context-aware application that helps consumers specify their information requirements with adaptive choice prompts in order to obtain desired supply context information.

Internal Context: The physical scope of a consumer context comprised of sensible body conditions that may influence the consumer's physiological needs. Certain context-aware applications can use bodily-attached sensors to keep track of the internal context information of mobile consumers.

Need Context: The conceptual part of a consumer context composed of stimuli that can influence the consumer's needs. A subset of need context that can be utilized by need context-aware applications is emergency context, from which the applications can infer the physiological and safety needs of consumers and provide services accordingly.

Proximate Context: The physical scope of a consumer context that is external to the body of consumer but close enough to be directly sensible to the consumer. Mobile consumers can describe and even record the information about their proximate contexts and share it with others.

Supply Context: The conceptual part of a consumer context composed of resources that can potentially supply what the consumer needs. Supply context-aware applications mainly help consumers obtain interesting and useful supply context information regarding their consumption needs.

Wireless Emergency Service (WES): A non-interactive mode of context-aware applications that provide necessary services in response to emergency contexts. Corresponding to the internal and external need contexts of mobile consumers, there are two types of WES: personal WES and public WES.

Wireless Local Community (WLC): A community mode of context-aware applications that facilitate the exchange of context information for a group of mobile consumers in a common supply context.

Core Principles of Educational Multimedia

Geraldine Torrisi-Steele
Griffith University, Australia

INTRODUCTION

The notion of using technology for educational purposes is not new. In fact, it can be traced back to the early 1900s during which school museums were used to distribute portable exhibits. This was the beginning of the visual education movement that persisted throughout the 1930s, as advances in technology such as radio and sound motion pictures continued. The training needs of World War II stimulated serious growth in the audiovisual instruction movement. Instructional television arrived in the 1950s but had little impact, due mainly to the expense of installing and maintaining systems. The advent of computers in the 1950s laid the foundation for CAI (computer assisted instruction) through the 1960s and 1970s. However, it wasn't until the 1980s that computers began to make a major impact on education (Reiser, 2001). Early applications of computer resources included the use of primitive simulation. These early simulations had little graphic capabilities and did little to enhance the learning experience (Munro, 2000).

Since the 1990s, there have been rapid advances in computer technologies in the area of multimedia production tools, delivery, and storage devices. Throughout the 1990s, numerous CD-ROM educational multimedia software was produced and was used in educational settings. More recently, the advent of the World Wide Web (WWW) and associated information and communications technologies (ICT) has opened a vast array of possibilities for the use of multimedia technologies to enrich the learning environment. Today, educational institutions are investing considerable effort and money into the use of multimedia. The use of multimedia technologies in educational institutions is seen as necessary for keeping education relevant to the 21st century (Selwyn & Gordard, 2003).

The term *multimedia* as used in this article refers to any technologies that make possible "the entirely digital delivery of content presented by using an integrated combination of audio, video, images (two-dimensional, three-dimensional) and text," along with the capacity to support user interaction (Torrisi-Steele, 2004, p. 24). Multimedia encompasses related communications technologies such as e-mail, chat, video-conferencing, and so forth. "The concept of interaction may be conceptualised as occurring along two dimensions: the capacity of the system to allow individual to control the pace of presentation and to make choices about which pathways are followed to move through the content; and the ability of the system to accept input from the user and provide appropriate feedback to that input.… Multimedia may be delivered on computer via CD-ROM, DVD, via the internet or on other devices such as mobile phones and personal digital assistants or any digital device capable of supporting interactive and integrated delivery of digital audio, video, image and text data" (Torrisi-Steele, 2004, p. 24).

The fundamental belief underlying this article is that the goal of implementing multimedia into educational contexts is to exploit the attributes of multimedia technologies in order to support deeper, more meaningful learner-centered learning. Furthermore, if multimedia is integrated effectively into educational contexts, then teaching and learning practice must necessarily be transformed (Torrisi-Steele, 2004). It is intended that this article will serve as a useful starting point for educators beginning to use multimedia. This article attempts to provide an overview of concepts related to the effective application of multimedia technologies to educational contexts. First, constructivist perspective is discussed as the accepted framework for the design of multimedia learning environments. Following this, the characteristics of constructivist multimedia learning environments are noted, and then some important professional development issues are highlighted.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

THEORETICAL FOUNDATIONS FOR THE ROLE OF MULTIMEDIA IN EDUCATIONAL CONTEXTS

Traditionally, teaching practices have focused on knowledge acquisition, direct instruction, and the recall of facts and procedures. This approach suited the needs of a society needing "assembly line workers" (Reigeluth, 1999, p. 18). However, in today's knowledge-based society, there is a necessity to emphasize deeper learning that occurs through creative thinking, problem solving, analysis, and evaluation, rather than the simple recall of facts and procedures emphasized in more traditional approaches (Bates, 2000). The advent of multimedia technologies has been heralded by educators as having the capacity to facilitate the required shift away from traditional teaching practices in order to innovate and improve on traditional practices (LeFoe, 1998; Relan & Gillani, 1997). Theoretically, the shift away from traditional teaching practices is conceptualized as a shift from a teacher-centered instructivist perspective to a learner-centered constructivist perspective on teaching and learning.

The constructivist perspective is widely accepted as the framework for design of educational multimedia applications (Strommen, 1999). The constructivist perspective describes a "theory of development whereby learners build their own knowledge by constructing mental models, or schemas, based on their own experiences" (Tse-Kian, 2003, p. 295). The constructivist view embodies notions that are in direct opposition to the traditional instructivist teaching methods that have been used in educational institutions for decades (see Table 1).

Expanding on Table 1, learning environments designed on constructivist principles tend to result in open-ended learning environments in which:

- Learners have different preferences of learning styles, cognitive abilities, and prior knowledge; they construct knowledge in individual ways by choosing their own pathways. Learning is affected by its contexts as well as the beliefs and attitudes of the learner;
- Optimal learning occurs when learners are active learners (e.g., learn by doing and learn by discovery;
- Learning is a process of construction whereby learners build knowledge through a process of scaffolding. Scaffolding is the process whereby learners link new knowledge with existing knowledge;
- Knowledge construction is facilitated through authentic problem-solving experiences;
- The process of learning is just as important as learning outcomes. Learners are encouraged to "articulate what they are doing in the environment and reasons for their actions" (Jonassen, 1999, p. 217).

Table 1. Key principles of the constructivist view of teaching and learning vs. key principles of the instructivist view of teaching and learning

CONSTRUCTIVIST	INSTRUCTIVIST
• learner-centered perspective: the learner is the focus of the learning environment – learners as individuals	• teacher-centered perspective: the teacher is focus of the learning environment- group learning
• encourages student independence in learning	• encourages student dependence on teacher
• teacher as facilitator that acts as a guide	• teacher as instructor
• learner and facilitator engage in a collaborative learning experience	• teacher in control of learning and in position of power over learner
• learners actively constructing knowledge in their own individual manner	• learners passively acquiring knowledge from the instructor
• Process of knowledge acquisition is important - how are learners interacting with the learning environment?	• acquisition of content and factual knowledge is key objective of learning episode
• curriculum design as development of knowledge spaces which allow active exploration by the learner	• curriculum design as goal oriented, strictly structured and ordered knowledge transmission
• Higher order thinking skills emphasized, creative thinking, problem solving, evaluation, synthesis	• behavioral objectives focusing on recall of facts and procedures, surface learning
• Open-ended learning environments (**OELE**)	• **directed instruction**

Multimedia, by virtue of its capacity f or interactivity, media integration, and communication, can be easily implemented as a tool for information gathering, communication, and knowledge construction. Multimedia lends itself well to the "creation and maintenance of learning environments which scaffold the personal and social construction of knowledge" (Richards & Nason, 1999). It is worth noting that the interactivity attribute of multimedia is considered extremely important from a constructivist perspective. Interactivity in terms of navigation allows learners to take responsibility for the pathways they follow in following learning goals. This supports the constructivist principles of personal construction of knowledge, learning by discovery, and emphasis on process and learner control. Interactivity in terms of feedback to user input into the system (e.g., responses to quizzes, etc.) allows for guided support of the learner. This also is congruent with constructivist principles of instruction as facilitation and also consistent with the notion of scaffolding, whereby learners are encouraged to link new to existing knowledge.

Using the constructivist views as a foundation, the key potentials of multimedia to facilitate constructivist learning are summarized by Kramer and Schmidt (2001) as:

- Cognitive flexibility through different accesses for the same topic;
- Multi-modal presentations to assist understanding, especially for learners with differing learning styles;
- "Flexible navigation" to allow learners to explore "networked information at their own pace" and to provide rigid guidance, if required;
- "Interaction facilities provide learners with opportunities for experimentation, context-dependent feedback, and constructive problem solving";
- Asynchronous and synchronous communication and collaboration facilities to bridge geographical distances; and
- Virtual laboratories and environments can offer near authentic situations for experimentation and problem solving.

THE EFFECTIVE IMPLEMENTATION OF MULTIMEDIA IN EDUCATIONAL CONTEXTS

Instructional Design Principles

Founded on constructivist principles, Savery and Duffy (1996) propose eight constructivist principles useful for guiding the instructional design of multimedia learning environments:

- Anchor all learning activities to a larger task or problem.
- Support learning in developing ownership for the overall problem or task.
- Design an authentic task.
- Design the tasks and learning environment to reflect the complexity of the environment that students should be able to function in at the end of learning.
- Give the learner ownership of the process to develop a solution.
- Design the learning environment to support and challenge the learner's thinking.
- Encourage testing ideas against alternative views and contexts.
- Provide opportunity for and support reflection on both the content learned and the process itself.

Along similar lines, Jonassen (1994) summarizes the basic tenets of the constructivist-guided instructional design models to develop learning environments that:

- Provide multiple representations of reality;
- Represent the natural complexity of the real world;
- Focus on knowledge construction, not reproduction;
- Present authentic tasks (contextualizing rather than abstracting instruction);
- Provide real-world, case-based learning environments rather than pre-determined instructional sequences;

- Foster reflective practice;
- Enable context-dependent and content-dependent knowledge construction; and support collaborative construction of knowledge through social negotiation, not competition among learners for recognition.

Professional Development Issues

While multimedia is perceived as having the potential to reshape teaching practice, oftentimes the attributes of multimedia technologies are not exploited effectively in order to maximize and create new learning opportunities, resulting in little impact on the learning environment. At the crux of this issue is the failure of educators to effectively integrate the multimedia technologies into the learning context.

[S]imply thinking up clever ways to use computers in traditional courses [relegates] technology to a secondary, supplemental role that fails to capitalise on its most potent strengths. (Strommen, 1999, p. 2)

The use of information technology has the potential to radically change what happens in higher education...every tutor who uses it in more than a superficial way will need to re-examine his or her approach to teaching and learning and adopt new strategies. (Tearle, Dillon, & Davis, 1999, p. 10)

Two key principles should underlie professional development efforts aimed at facilitating the effective integration of technology in such a way so as to produce positive innovative changes in practice:

Principle 1: Transformation in practice as an evolutionary process

Transformation of practice through the integration of multimedia is a process occurring over time that is best conceptualized perhaps by the continuum of stages of instructional evolution presented by Sandholtz, Ringstaff, and Dwyer (1997):

- **Stage One:** Entry point for technology use where there is an awareness of possibilities, but the technology does not significantly impact on practice.
- **Stage Two:** Adaptation stage where there is some evidence of integrating technology into existing practice
- **Stage Three:** Transformation stage where the technology is a catalyst for significant changes in practice.

The idea of progressive technology adoption is supported by others. For example, Goddard (2002) recognizes five stages of progression:

- **Knowledge Stage:** Awareness of technology existence.
- **Persuasion Stage:** Technology as support for traditional productivity rather than curriculum related.
- **Decision Stage:** Acceptance or rejection of technology for curriculum use (acceptance leading to supplemental uses).
- **Implementation Stage:** Recognition that technology can help achieve some curriculum goals.
- **Confirmation Stage:** Use of technology leads to redefinition of learning environment—true integration leading to change.

The recognition that technology integration is an evolutionary process precipitates the second key principle that should underlie professional development programs—reflective practice.

Principle 2: Transformation is necessarily fueled by reflective practice

A lack of reflection often leads to perpetuation of traditional teaching methods that may be inappropriate and thus fail to bring about "high quality student learning" (Ballantyne, Bain & Packer, 1999, p. 237). It is important that professional development programs focus on sustained reflection on practice from the beginning of endeavors in multimedia materials development through completion stages, followed by debriefing and further reflection feedback into a cycle of continuous evolution of thought and practice. The need for educators to reflect on their practice in order to facilitate effective and transformative integration of multimedia technologies cannot be understated.

In addition to these two principles, the following considerations for professional development programs, arising from the authors' investigation into the training needs for educators developing multimedia materials, are also important:

- The knowledge-delivery view of online technologies must be challenged, as it merely replicates teacher-centered models of knowledge transmission and has little value in reshaping practice;
- Empathising with and addressing concerns that arise from educators' attempts at innovation through technology;
- Equipping educators with knowledge about the potential of the new technologies (i.e., online) must occur within the context of the total curriculum rather than in isolation of the academic's curriculum needs;
- Fostering a team-orientated, collaborative, and supportive approach to online materials production;
- Providing opportunities for developing basic computer competencies necessary for developing confidence in using technology as a normal part of teaching activities.

LOOKING TO THE FUTURE

Undeniably, rapid changes in technologies available for implementation in learning contexts will persist. There is no doubt that emerging technologies will offer a greater array of possibilities for enhancing learning. Simply implementing new technologies in ways that replicate traditional teaching strategies is counterproductive. Thus, there is an urgent and continuing need for ongoing research into how to best exploit the attributes of emerging technologies to further enhance the quality of teaching and learning environments so as to facilitate development of lifelong learners, who are adequately equipped to participate in society.

CONCLUSION

This article has reviewed core principles of the constructivist view of learning, the accepted framework for guiding the design of technology-based learning environments. Special note was made of the importance of interactivity to support constructivist principles. Design guidelines based on constructivist principles also were noted. Finally, the importance of professional development for educators that focuses on reflective practice and evolutionary approach to practice transformation was discussed. In implementing future technologies in educational contexts, the goal must remain to improve the quality of teaching and learning.

REFERENCES

Ballantyne, R., Bain, J.D., & Packer, J. (1999). Researching university teaching in Australia: Themes and issues in academics' reflections. *Studies in Higher Education, 24*(2), 237-257.

Bates, A.W. (2000). *Managing technological change*. San Francisco: Jossey-Bass.

Goddard, M. (2002). What do we do with these computers? Reflections on technology in the classroom. *Journal of Research on Technology in Education, 35*(1), 19-26.

Hannafin, M., Land, S., & Oliver, K. (1999). Open learning environments: Foundations, methods and models. In C. Reigeluth (Ed.), *Instructional-design theories and models* (pp. 115-140). Hillsdale, NJ: Erlbaum.

Jonassen, D.H. (1994). Thinking technology: Toward a constructivist design model. *Educational Technology, Research and Development, 34*(4), 34-37.

Jonassen, D.H. (1999). Designing constructivist learning environments. In C. Reigeluth (Ed.), *Instructional-design theories and models* (pp. 215-239). Hillsdale, NJ: Erlbaum.

Kramer, B.J., & Schmidt, H. (2001). Components and tools for on-line education. *European Journal of Education*, 36(2), 195-222.

Lefoe, G. (1998). *Creating constructivist learning environments on the Web: The challenge of higher education.* Retrieved August 10, 2004, from *http://*

www.ascilite.org.au/conferences/wollongong98/ascpapers98.html

Munro, R. (2000). Exploring and explaining the past: ICT and history. *Educational Media International, 37*(4), 251-256.

Reigeluth, C. (1999). What is instructional-design theory and how is it changing? In C. Reigeluth (Ed.), *Instructional-design theories and models* (pp. 5-29). Hillsdale, NJ: Erlbaum.

Reiser, R.A. (2001). A history of instructional design and technology: Part I: A history of instructional media. *Educational Technology, Research and Development, 49*(1), 53-75.

Relan, A., & Gillani, B. (1997). Web-based instruction and the traditional classroom: Similarities and differences. In B.H. Khan (Ed.), *Web-based instruction* (pp. 41-46). Englewood Cliffs, NJ: Educational Technology Publications.

Richards, C., & Nason, R. (1999). Prerequisite principles for integrating (not just tacking-on) new technologies in the curricula of tertiary education large classes. In J. Winn (Ed.) *ASCILITE '99 Responding to diversity conference proceedings*. Brisbane: QUT. Retrieved March 9, 2005 from *http://www.ascilite.org.au/conferences/brisbane99/papers/papers.htm*

Sandholtz, J., Ringstaff, C., & Dwyer, D. (1997). *Teaching with technology.* New York: Teachers College Press.

Savery J.R. & Duffy T.M. (1996). An instructional model and its constructivist framework. In B Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design.* Englewood Cliffs, NJ: Educational Technology Publications.

Selwyn, N., & Gorard, S. (2003). Reality bytes: Examining the rhetoric of widening educational participation via ICT. *British Journal of Educational Technology, 34*(2), 169-181.

Strommen, D. (1999). *Constructivism, technology, and the future of classroom learning.* Retrieved September 27, 1999, from *http://www.ilt.columbia.edu/ilt/papers/construct.html*

Tearle, P., Dillon, P., & Davis, N. (1999). Use of information technology by English university teachers. Developments and trends at the time of the national inquiry into higher education. *Journal of Further and Higher Education, 23*(1), 5-15.

Torrisi, G., & Davis, G. (2000). Online learning as a catalyst for reshaping practice—The experiences of some academics developing online materials. *International Journal of Academic Development, 5*(2), 166-176.

Torrisi-Steele, G. (2004). Toward effective use of multimedia technologies in education In S. Mishra & R.C. Sharma (Eds.), *Interactive multimedia in education and training* (pp. 25-46). Hershey, PA: Idea Group Publishing.

Tse-Kian, K.N. (2003). Using multimedia in a constructivist learning environment in the Malaysian classroom. *Australian Journal of Educational Technology, 19*(3), 293-310.

KEY TERMS

Active Learning: A key concept within the constructivist perspective on learning that perceives learners as mentally active in seeking to make meaning.

Constructivist Perspective: A perspective on learning that places emphasis on learners as building their own internal and individual representation of knowledge.

Directed Instruction: A learning environment characterized by directed instruction is one in which the emphasis is on "external engineering" (by the teacher) of "what is to be learned" as well as strategies for "how it will be learned" (Hannafin, Land & Oliver, 1999, p. 122).

Instructivist Perspective: A perspective on learning that places emphasis on the teacher in the role of an instructor that is in control of what is to be learned and how it is to be learned. The learner is the passive recipient of knowledge. Often referred to as teacher-centered learning environment.

Interactivity: The ability of a multimedia system to respond to user input. The interactivity element of multimedia is considered of central importance from the point of view that it facilitates the

active knowledge construction by enabling learners to make decisions about pathways they will follow through content.

Multimedia: The entirely digital delivery of content presented by using an integrated combination of audio, video, images (two-dimensional, three-dimensional) and text, along with the capacity to support user interaction (Torrisi-Steele, 2004).

OELE: Multimedia learning environments based on constructivist principles tend to be open-ended learning environments (OELEs). OELEs are open-ended in that they allow the individual learner some degree of control in establishing learning goals and/or pathways chosen to achieve learning.

Reflective Practice: Refers to the notion that educators need to think continuously about and evaluate the effectiveness of the strategies and learning environment designs they are using.

Corporate Conferencing

Vilas D. Nandavadekar
University of Pune, India

INTRODUCTION

Today's corporate need for manpower is growing—the number of remote relationships, mobile workers, and virtual teams. The efficiency and effectiveness of manpower is real success of the corporation, which largely depends on collaborative work. The difficulty faced by the organization is in the scheduling and execution of meetings, conferences, and other events.

The work becomes easier and simpler by using Corporate Conferencing (CC) today. Corporate Conferencing is used in the delivery, control, and execution of scheduling of work/event effectively. It optimizes conferencing services quality and costs by increasing an organization's flexibility to deliver services that suit the end user/customer needs. It removes obstacles between organization and virtual teams. It keeps track of mobile workers by improving accessibility of conferencing technologies. It enhances facilities and organizations' capabilities by providing corporate conferencing. It reduces capital cost of administration. It improves utilization and conferencing space and resources like 3Ps (People, Process, and Problem).

BACKGROUND

As more and more organizations compete globally and/or rely on suppliers throughout the world, the business need for enhanced communications capabilities and higher availability mounts steadily. The third major driving force for the movement to interactive corporate communications is the need for additional and more frequent collaboration. There cannot be a better two-way communication system for a group of users across a small geography (Saraogi, 2003). Many organizations are finding that collaborating using interactive devices, along with document sharing, streamlines their business activities by decreasing time to market and by increasing productivity. Meanwhile, reductions in business travel since the tragedies of September 11, 2001, are placing more demands on corporate conferencing to manage 3Ps. If education is conceived as a way of changing students, then educators should accept that they cannot be culturally benign, but invariably promote certain ways of being over others (Christopher, 2001).

Based on data of Wainhouse Research, it determined that almost two-thirds (64%) of business travelers considered access to audio, video, and Web conferencing technologies to be important to them in a post-work environment. The World Wide Web, fax, video, and e-mail enable the quick dissemination of information and immediate communication worldwide. The inclusion of women will require a concerted effort to overcome the gender bias and stereotypes that have haunted those wanting to become involved in aspects of the field on a managerial level, such as conferencing.

Certainly, teaching in an online environment is influenced by the absence of the non-verbal communication that occurs in the face-to-face settings of conventional education, and the reduction in the amount of paralinguistic information transmitted, as compared to some other modes of distance education such as video or audio teleconferencing (Terry, 2001). To attend meetings personally is very important for the effective performance of business today. But attending in person is not always possible. There are several reasons for this, most of which are:

1. **Time:** To travel long distance and attend meeting is very difficult.
2. **Cost:** The cost of the travel for attending meeting personally.
3. **Workload:** Difficult to attend because of some other work/duty.
4. **Stress:** Too much stress on employees/staff.
5. **Decision:** Too much delay in decision making.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

METHODS OF CONFERENCING

To overcome these problems, we can better choose one of the methods of corporate conferencing. These methods are as follows:

Video Conferencing

It delivers and provides live session in true fashion in the world. Video conferencing allows a face-to-face meeting to take place between two or more geographical locations simultaneously. This is the advantage over an audio conference or normal telephone call. In this method, we can observe performance as well as reaction of people. It is possible to take decision in time. It also defines to engage communication and transmission between two or more persons/parties in different geographical locations via video and audio through a private network or Internet. It allows face-to-face conversations.

Video conferencing means greatly increased bandwidth requirements. It requires high bandwidth. This is one of the drawbacks of this method. Video is somewhat complex to access, as there are several choices to be made. Required bandwidth is massively influenced by the size of the video image and the quality. Quality is determined by the compression rate (how good is the image) and the update rate (how many images are displayed per second). Typically, video conferencing requires between 200kb/s and 1,000 kb/s per user. Please note that this means neither full screen nor TV quality video. The implication is that even small and not very fluent video requires significant bandwidth, both at the user's end and even more at the server's. Large groups require a dedicated broadband network (Wilheim, 2004).

TV companies typically compress to around 24 Mbps to 32 Mbps. However, this still results in higher transmission costs that would normally be acceptable for any other business. The coder takes the video and audio and compresses them to a bit stream that can be transmitted digitally to the distant end. With improved coding techniques, the bit stream can be as low as 56 kbps, or up to 2 Mbps. For business quality conferencing, 384 kbps is the industry standard. The decoder extracts the video and audio signals from the received bit stream and allows for the signal to be displayed on a TV and heard through the speakers. In addition to video and audio, user data can be transmitted simultaneously to allow for the transfer of computer files, or to enable users to work collaboratively on documents. This latter area has become increasingly important with the availability of effective data collaboration software (e.g., from entry level to performance, Polycom Group Video Conferencing Systems offers a wide range of choices to suit any application environment, from the office to the board room, the courtoom to the classroom).

Web Conferencing

Web-based collaboration offers definite benefits: it is easy, it is cost-effective, and it allows companies to do multiple activities in a seamless fashion. But virtual teams are not without disadvantage. For one thing, virtual teams must function with less direct interaction among members. So, virtual team members require excellent project management skills, strong time management skills, and interpersonal awareness. In addition, they must be able to use electronic communication and collaboration technologies, and they need to be able to work across cultures (Bovee, 2004). A communication is conducted via the WWW between two or more parties/persons in different geographical locations. It is in the form of synchronous real time or in an asynchronous environment (at our convenience and our own time).

Web casting allows greater access to significantly extend the reach of the meeting, far beyond the attendees to a much wider audience. The event was Web cast live and is also available for on-demand viewing, enabling the employees/public to view at their convenience (Greater, 2004). Furthermore, recent research has shown that an overlaid network may cost up to 20% less to operate, compared to deploying rule-based (Internet protocol) communications internally over the corporate network (WAN) (Brent, 2002). Traditional video conferencing solutions tend to be overly expensive and very bandwidth hungry. Existing Web conferencing solutions lack rich media support and shared applications (e.g., MeetingServer is a carrier-grade, high-function, Web conference server solution that allows service providers to deploy a robust, scalable, manageable Web conferencing service to consumers, enterprises, and virtual ISPs.

Computer Conferencing

The online conferencing model enhances traditional methods in five ways: (1) text-based: forces people to focus on the message, not the messenger; makes thinking tangible; and forces attentiveness; (2) asynchronous: the 24-hour classroom is always open; plenty of time for reflection, analysis, and composition; encourages thinking and retrospective analysis; the whole transcript discussion is there for review; class discussion is open ended, not limited to the end of period; (3) many-to-many: learning groups of peers facilitate active learning, reduce anxiety, assist understanding, and facilitate cognitive development; and resolve conceptual conflict in the groups; (4) computer mediated: encourages active involvement, as opposed to the passive learning from books or lectures; gives learner and teacher control; interactions are revisable, archivable, and retrievable; hypermedia tools aid in structuring, interconnecting, and integrating new ideas; and (5) place independent: not constrained by geography; panoptic power; collaboration with global experts online and access to global archival resources; access for the educationally disenfranchised (Barry 2003).

Computer conferencing is exchanging information and ideas such as in multi-user environments through computers (e.g., e-mail). Computer conferencing can impose intellectual rigor; it can be the premier environment for writing through the curriculum and one of the best ways to promote active, student-centered learning (Klemm). For example, Interactive Conferencing Solutions EasyLink delivers a complete range of audio conferencing and Web conferencing solutions. We connect thousands of business professionals around the globe every day, and we know that success comes from focusing on one call at a time. The end result—reliable, easy-to-use Internet conferencing services that are perfectly tailored to meet your business communication needs.

Present Conferencing

Web services provide organizations with a flexible, standards-based mechanism for deploying business logic and functionality to distributed people. There are different tools available in the market today. Some of these tools are as follows:

In traditional methods of scheduling they use:

- **Manual Scheduling Method:** In this method, people plan their work according to the schedule and records, and otherwise schedule their work with other resources. In this method, they use handwritten notice, paper, phone calls, chatrooms or e-mail messages and personal information (i.e., through Palm). This method is inefficient, unscalable, and difficult to mange by people within or outside the organization.
- **Calendaring and Group Messaging:** In this method, scheduling can be done by using group messaging and calendaring. In this method, they use any ready-made calendar like an Oracle calendar, Lotus Notes, or Microsoft Outlook and do group messaging to all participants or workers. Calendaring and group messaging requires high integration, distribution, and control over the 3Ps.
- **Collaborative and Specialized Service Scheduling:** In this method, they use ready-made software like Web conferencing service scheduler. This is more suitable for all the middle- as well as large-scale organizations for organizing conferences. It is unified, managed, and distributed scheduling of all conferencing activities in the corporate environment. A collaborative effort must be in place to ensure that everyone gets the information most relevant to them (Weiser, 2004).

USAGE OF CORPORATE CONFERENCING FOR 3PS

Conferencing is a necessary complement to the communications capabilities of any competitive business in the 21st century. With the help of video to IP (Internet protocol) and personal computer/laptop computers, cost-effectiveness has brought corporate conferencing within the reach of practically any business. With the help of less manpower (people), we can organize and plan quality conferences. Today's processes (technology) help us to be able to do desktop conferencing instead of doing meetings in meeting rooms. Most of the industry uses modern presentation styles or discussion rooms (i.e., PowerPoint presentations) for better understanding and communication. Corporate conferencing is play-

ing a vital role in many meetings today. Whether it is a Fortune 500 company or a small to medium player in corporate, the age of video conferencing has become an integral part of day-to-day success (www.acutus.com).

The process is important for corporate conferencing, which is at the top of the list as a necessary tool for corporate communications. The equipment required for corporate conferencing is easy to install, network-friendly, easy to operate (i.e., a computer, telephone set, etc.), and has better quality outputs by using TV/CD/VCR. The speed of data retrieval and data transfer is very high, and it is available at low cost. These equipments are best suited for the corporation to perform or organize conferences. Corporate conferencing refers to the ability to deliver and make schedules of all events of the meeting, conference, or other collective work in a unified, manageable fashion. The real-world applications for conference calls are limitless. Students, teachers, employees, and management can and should be benefiting from this exciting, convenient technology (www.conference-call-review.com).

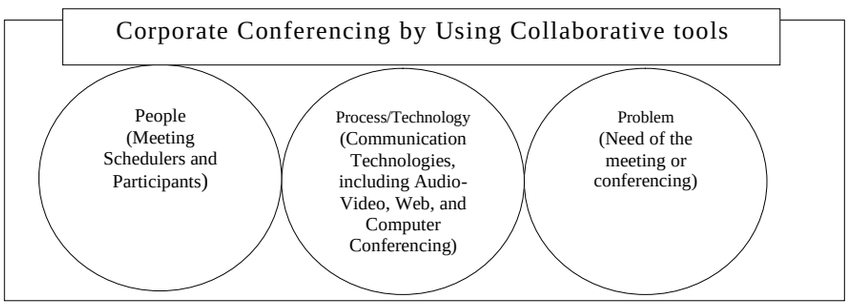
CORPORATE CONFERENCING: WHAT TO LOOK FOR?

Corporate conferencing helps to change/modify business process with the help of ready-made software. The following table describes benefits of corporate conferencing.

Table 1. Benefits of corporate conferencing in terms of payback and other factors

Sr. No	Who will get returns of CC?	What type of benefit do they have?
1.	Management/ Executives	a) Workload Sharing – labor saving. CC not only help utilize physical resources better, but it reduces the labor cost. b) To decide and frame policy for organization. c) Centralized service for decentralized workers or for virtual teams. d) Less infrastructure and reusable forms of resources.
2.	Employees/ workers: Job inside or outside/onsite of the organization	a) One stops scheduling for employees. It synchronizes scheduling and prevents conflicts from occurring with other employees. b) It improves efficiency of workers because of universal access. It provides the ability to plan, accept, invite, and extend conferences from anywhere at any point of time. Without any disruptions or interruptions, it clarifies uncertainties when in scheduling/planning mode. It is more useful to increase productivity of organizations without spending much time on rescheduling.
3.	Departments like EDP, Information Technology, etc.	a) It generates revenue for the department in the form of development. b) It helps the virtual teams in IT departments when they are working onsite on a project.
4.	Organization	a) It helps to keep track of virtual teams. b) It helps organizations to make quick decisions in uncertainty state. c) Soft and hard saving: To make payback to an organization has to do with the rate of return of the corporate conferencing as measured in both hard savings and soft savings. The hard savings are those areas that can be measured in terms of manpower. It also includes savings based on utilization, and unnecessary costs can be eliminated. Soft savings also can be found in organizations that are self-serve already in their approach to meeting management. These savings can be based on the delivery of a platform that enriches the scheduling experience, while keeping it simplified and convenient. A meeting without wasting time. It has an impact on the productivity of the organization.
5.	Staff: who works for CC as a service staff	It requires very less staff for operation. In manual conferencing, we need five persons per month, five5 days a week, for four weeks, at eight hours per day (800 total hours for manual scheduling). But in the case of corporate conferencing, we need only one person per month, who will work as the administrator. Total work hours for CC is one person per month, five days a week, for four weeks, at eight hours per day, or 160 hours. Automatically, we require less service staff for CC as compared to other conferencing methods.

Figure 1. List of items for collaboration and integration of corporate conferencing


Corporate Conferencing by Using Collaborative tools
People (Meeting Schedulers and Participants)
Process/Technology (Communication Technologies, including Audio-Video, Web, and Computer Conferencing)
Problem (Need of the meeting or conferencing)

3PS: THE COMPLEXITY OF MANAGING CORPORATE CONFERENCING

In an organization, the task/work of managing and maintaining 3Ps for resources never ends. It holds real joint hands and collaboration with each other. Figure 1 indicates items that are important for collaboration and integration of corporate conferencing.

The list looks simple (Figure 1), but the ability to bring together and manage disparate items is far from simple. It directly impacts the ability to effectively and efficiently corporate conference. These items are the least items. It provides different needs at different times for conferencing.

CONCLUSION

In today's world, most organizations are applying corporate conferencing method for scheduling their work, meetings, and so forth. They conduct their meetings through Web, video, or computer via a network or the Internet. It has a greater flexibility in terms of space, which includes greater than ever meetings. In this, the end user takes the benefits, which will result in the experience of low operation costs in terms of 3Ps (people, process, and problem). It is more transparent and has smarter capabilities. It is useful for management for better output results. It drives business intelligence and analytics for understanding. It improves efficiencies, maximizes productivity, and increases profits. The real-world applications for corporate conferencing are limitless. All the people (learners) (e.g., students, teachers, employees, and management) can and should be benefiting from this exciting, convenient technology.

REFERENCES

Anderson, T., & Liam, R.D. (2001). Assessing teaching presence in a computer conferencing context. *Journal of Asynchronous Learning Networks, 5*(2).

Bovee, Thill, and Schatzman. (2004). *Business communication today*. Singapore: Pearson Education.

Fubini, F. (2004). *Greater London authority selects virtue to deliver public meeting*. London: Virtue Communications.

Harrington, H., & Quinn-Leering, K. (1995). Reflection, dialogue, and computer conferencing. *Proceedings of the Annual Meeting of the American Educational Research Association*, San Francisco.

http://www.acutus.com/corporate.asp/

http://www.acutus.com/presentation/presentation_files/slide0132.htm

http://www.conference-call-review.com/

http://search.researchpapers.net/cgi-bin/query?mss=researchpapers&q=Video%20Conferencing

http://www.wainhouse.com/files/papers/wr-converged-networking.pdf

Kelly, B.E. (2002). IP telecommunications: Rich media conferencing through converged networking. *Infocomm*.

Klemm, W.R., & Snell, J.R. Instructional design principles for teaching in computer conferencing environments. *Proceedings of the Distance Education Conference, Bridging Research and Practice*, San Antonio, Texas.

Lea, M.R. (1998). Academic literacies and learning through computer conferencing. *Proceedings of Higher Education Close Up*, University of Central Lancashire, Preston.

Prashant, S., & Sanjay, S. (2003). Radio trunking services: Bulky and beautiful. *Voice and Data — The Business Communication, 9*(9).

Shell, B. (2003). *Why computer conferencing?* British Columbia, Canada: The Centre For Systems Science, Simon Fraser University.

Wainhouse Research. (2002). Conferencing technology and travel behaviour. Web conferencing [technical white paper]. (2004). Retrieved from *www.virtue-communications.com*

Weiser, J. (2004). Quality management for Web services—The requirement of interconnected business. *Web Services Journal, SYS-CON Media, Inc.*

Wilheim & Muncih. (2004). *Web conferencing* [technical white paper]. London: Virtue Communications.

Ziguras, C. (2001). Educational technology in transnational higher education in South East Asia: The cultural politics of flexible learning. *Educational Technology & Society, 4*(4), 15.

KEY TERMS

Asynchronous: The 24-hour classroom discussion is always open; plenty of time for reflection, analysis, and composition; encourages thinking, retrospective analysis; the whole transcript discussion is there for review; class discussion is open ended, not limited to the end of period.

Collaborative Tools: A set of tools and techniques that facilitate distant collaboration geographically at different locations.

Computer Conferencing: Exchanging information and ideas in a multi-user environment through computers (e.g., e-mail).

Corporate Communications: It is a broadcasting (provides organizations with the technology infrastructure and software solutions that empower them to create and deliver communication messages) leading corporate communications solutions provider in enabling corporate to communicate both internally among employees and externally (out side the organization) to support their business needs and goals; operationally less costly.

IP: Internet Protocol is a unique number or address that is used for network routing into the computer machine.

Synchronous: To make event/meeting/discussion happen at the scheduled time at different locations for different groups of people. It is basically used to create face-to-face environments.

Video Conferencing: Engage communication and transmission between two or more persons/parties in different geographical locations via video and audio through a private network or Internet. It allows face-to-face conversations.

Web Cast: Communications between one or many persons through electronic media. A communication made on the World Wide Web.

Web Conferencing: A communication conducted via the WWW between two or more parties/persons in different geographical locations. It is in the form of synchronous real time or in an asynchronous environment (at your convenience).

Cost Models for Telecommunication Networks and Their Application to GSM Systems

Klaus D. Hackbarth
University of Cantabria, Spain

J. Antonio Portilla
University of Alcala, Spain

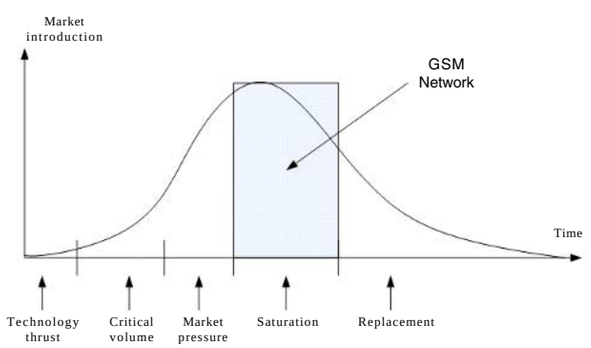
Ing. Carlos Diaz
University of Alcala, Spain

INTRODUCTION

Currently mobile networks are one of the key issues in the information society. The use of cellular phones has been broadly extended since the middle 1990s, in Europe mainly with the GSM (Global System for Mobile Communication) system, and in the United States (U.S.) with the IS-54 system. The technologies on which these systems are based, Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA) are completely developed, the networks are completely deployed and the business models are almost exhausted[1] (Garrese, 2003). Therefore, these systems are in the saturation stage if we consider the network life cycle described by Ward, which is shown in Figure 1.

At this stage, it is possible to assume that all work is over in this field. However, in this time stage there are new critical problems, mainly related with network interconnection, regulation pricing and accounting.

Figure 1. Network life cycle (Source Ward, 1991)



These types of questions are quite similar to the regulatory issues in fixed networks in the fields of Public Switched Telephone Network (PSTN), Integrated Service Data Network (ISDN) and Digital Subscriber Line (DSL) access. In the European environment, there is an important tradition in these regulatory issues, mainly produced by the extinction of the old state-dominant network operators and market liberalization. National Regulatory Authorities (NRAs) give priority to guarantee the free competition through different strategic policies that apply mainly to the following topics:

- **Interconnection and call termination prices:** The most common situation is a call originated in the network of operator A, and terminates in a customer of another network operator, B. There are other scenarios, like transit interconnection, where a call is originated and terminated in the network of operator A but has to be routed through the network of operator B. In any case, the first operator has to pay some charge to the second one for using its network. The establishment of a fair charge is one of the key points of regulatory policies.
- **Universal service tariffs:** In most countries, the state incumbent operator had a monopolistic advantage; hence, the prices were established by a mixture of historical costs and political issues. Currently, with market liberalization and the entry of new operators, these tariffs must be strictly observed to avoid unfair practices.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

- **Retail and wholesale services (customer access):** This situation deals mainly with the local loop; that is, the final access to the customer. An example is when a network operator offers physical access to the customer—the copper line in DSL access, and an Internet Service Provider (ISP) offers the Internet access.

The establishment of these prices, tariffs and other issues related with the regulatory activities requires defining cost methodologies to provide an objective framework.

The following sections present different cost methodologies applied in telecommunication networks. Furthermore, a specific model named *Forward-Looking Long-Run Incremental Cost (FL-LRIC)* is deeper studied. Finally, the FL-LRIC model is applied to the specific case of the GSM mobile network.

COST METHODOLOGIES

Cost methodologies must ensure that prices led to profitability, or that they at least cover the proper costs (cost-based prices). A fundamental difficulty in defining cost-based pricing is that different services usually use common network elements. A large part of the total cost is a common cost; hence, it is difficult to divide the different services. The cost-based prices must perform three conditions (Courcoubetis, 2003):

- **Subsidy free prices:** each customer has to pay only for its service.

Figure 2. Bottom-up approach

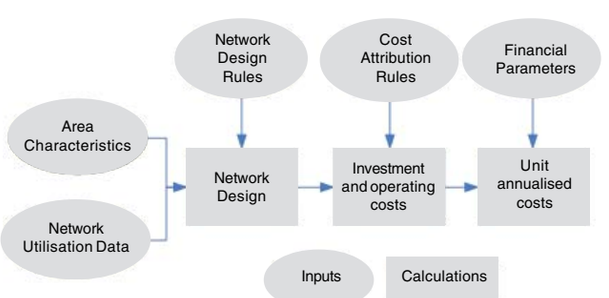
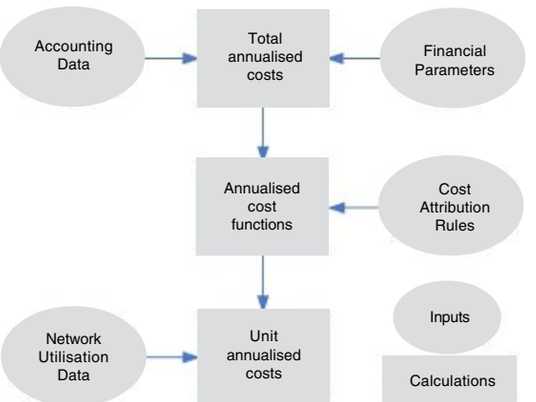


Figure 3. Top-down approach



- **Sustainable prices:** prices should be defensive against competition.
- **Welfare maximization:** prices should ensure the social welfare maximization.

Note that the three conditions could be mutually incompatible. The aim of welfare maximization may be in conflict with the others, restricting the feasible set of operating points. Several methods (Mitchell, 1991; Osborne, 1994) have been developed for the cost-based prices calculation, but they have practical restrictions; that is, the ignorance of complete cost functions. This article presents a set of practical methods for the calculation of the cost of services that fulfill the conditions mentioned.

In practice, the main problem is the distribution of common costs between services. Usually only a small part of the total cost is comprised of factors that can be attributed to a single service. The common costs are calculated, subtracting the cost imputable to each service to the total cost. There are two alternatives for the calculation of the common cost: top-down and bottom-up (see Figures 2 and 3, respectively).

In the bottom-up approach, each cost element is computed using a model of the most efficient facility specialized in the production of the single service, considering the most efficient current technology. Thus, we construct the individual cost building models of fictitious facilities that produce just one of these services. The top-down approach starts from

the given cost structure of existing facilities and attempts to allocate the cost that has actually incurred to the various services.

Additionally, according to Courcoubetis (2003), a division between direct and indirect costs and fixed and variable costs should be considered. Direct cost is the part solely attributed to a particular service and will cease to exist if the service is no longer produced. Indirect costs are related only to the provision of all services. Fixed costs is the value obtained by the addition of the costs independent of the service quantity. That means these costs remain constant when the quantity of the service changes. Opposite are variable costs, because they depend on the amount of the service produced.

Several methodologies calculate the price under the previous cost definition Most relevant are the two introduced below (Taschdjian, 2001):

- **Fully Distributed Cost (FDC):** The idea of FDC is to divide the total cost that the firm incurs amongst the services that it sells. This is a mechanical process; a program takes the values of the actual costs of the operating factors and computes for each service its portion. FDC is a top-down approach.
- **FL-LRIC:** This is a bottom-up approach, in which the costs of the services are computed using an optimized model of the network and service production technologies.

Table 1 shows the main advantages and disadvantages of these methods.

Currently, regulation studies are mainly based on the FL-LRIC (see European Commission, 1998).

FL-LRIC COST METHODOLOGY

The objective of the FL-LRIC cost model is to estimate the investment cost incurred by a new hypothetical entrant operator under particular conditions. This new operator has to provide the same service briefcase as the established one. Furthermore, the new operator has to define an optimal network configuration using the most suitable technology (Hackbarth, 2002).

Using the FL-LRIC methodology, market partners can estimate the price $p(A)$ of a corresponding service A. The underlying concepts to perform this estimation are introduced next.

The concept of *Forward Looking* implies performing the network design. It is considered both present and forecast future of customer demand. Furthermore, the *Long Run* concept means that we consider large increments of additional output, allowing the capital investment to vary.

The incremental cost of providing a specific service in a shared environment can be defined as the common cost of joint production subtracting the independent cost of the rest of the services. Therefore, if we consider two different services, A and B, the incremental cost of providing A service can be defined as

$$LRIC\ (A) = C(A,B)\text{-}C(B)$$

Table 1. Comparison between FDC and FL-LRIC methodologies

Costing method	Advantages	Disadvantages
FDC	- The full cost can be recovered - The cost computation process is easier than in other models.	- Prices may result unduly high - Adopting historical costs may induce wrong decisions in future
FL-LRIC	- The use of a prospective cost basis allows estimation of the expectations of competitive operators	- It does not allow for the full recovery of sustained money

Where C(A,B) is the joint cost of providing services A and B, and C(B) is the cost of providing service B independently. The methodology for implementing LRIC is based on constructing bottom-up models from which to compute C(A,B) and C(B), considering current costs[2].

Note that the sum of the service prices calculated under the LRIC model do not cover the costs of joint production, because the term [C(A,B)-C(A)-C(B)] is usually negative.

$$LRIC(A)+LRIC(B)= C(A,B) + [C(A,B)-C(A)-C(B)]$$

Therefore, the price of the service A, $p(A)$, has to be set between the incremental cost of the service LRIC(A) and the stand-alone cost C(A).

$$LRIC\ (A) \leq p(A) \leq C(A)$$

As previously mentioned, the LRIC requires a model and the corresponding procedure to estimate a realistic network design, allowing calculation of the network investment. The next section deals with the particular application of the model to GSM mobile networks, focusing on network design, dimensioning and the corresponding cost calculation.

FL-LRIC APPLIED TO GSM MOBILE NETWORKS

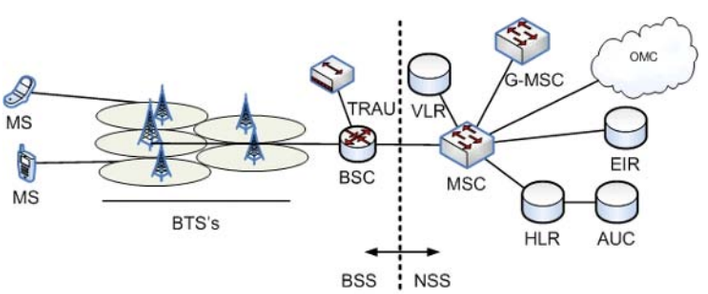
Contrary to fixed networks, the application of FL-LRIC cost models to mobile networks, and specifically to a GSM-PLMN (Public Land Mobile Network), has some particular features that have to be considered, due partly to the radio link-based network. The network design and configuration depends on several issues, such as general parameters of the operator (service briefcase, market share, coverage requirements, equipment provider), demographic and geographic parameters (population, type of terrain, building concentration) and so on. Obviously, a critical design parameter is the technology and network hierarchy. The reference architecture of a GSM network is shown in Figure 4.

Note that there are two main subsystems: the Base Station Subsystem (BSS), which corresponds to the access network; and the Network Switching Subsystem (NSS), which keeps with the conveyance network. Considering a design scope, the BSS can be further divided into a cell deployment level, which consists of the Base Station Tranceivers (BTS) and a fixed-part level, which corresponds to the Base Station Controllers (BSC) and Transcoding Rate Adaptation Units (TRAU).

The design of an optimal GSM-PLMN network on a national level, required for the bottom-up approach of the LRIC model, is a huge task. This is due to the number and complexity of heterogeneous planning scenarios, mainly in the cell deployment level (all the cities and municipalities of the country). Therefore, the complete set of scenarios must be reduced to a limited but representative one, and perform the design considering only a specific example for each type. Afterwards, the results have to be extrapolated to cover the national network. A possible set of scenarios with their mapping in the Spanish case are the following:

- Metropoly cities; for example, Madrid (5,719,000 inhabitants)

Figure 4. GSM network architecture



- Medium-size cities; for example, Zaragoza (601,674 inhabitants)
- Small cities and villages; for example, El Astillero (15,000 inhabitants)
- Roads, highways and railroads
- Countryside.

For each scenario, the number of BTS required to provide the corresponding quality of service (QoS) to the customers must be calculated. In this process, several factors have special relevance. The network planner requires detailed information about the different types of available BTS. After that, the cell radius has to be obtained through specific coverage and capacity studies. The coverage can be obtained using analytical methods (Okumura, 1968; Maciel, 1993; COST 231, 1991), providing a maximum value of the cell radius. Using this value and with the number of available channels on the selected BTS and traffic parameters (user call rate, connection time and customer density), the network planner can test if the target QoS is reached. For this purpose, a traffic model is required; the most relevant was developed by Rappaport (1986). If the QoS is not reached, several mechanisms can be used, where the most important are sectoring and with "umbrella cells"[3]. The amount of required BTS of each type are obtained by the division of the extension of each particular area of the city between the coverage areas of the BTS assigned to it. Additionally, the maximum number of cells is limited by the frequency reusing factor determined by the number of different frequency channels assigned to the operator. Further information about this topic can be found in Hernando-Rabanos (1999).

Remember that the objective of the network design in the LRIC model is to calculate the use factor of the different network elements by each unit of user traffic. To obtain this use factor, a projection model is defined. Taking the cell as the reference level, we have to find the use factor of the different network elements by each type of cell. (Note that each type of cell is defined by the type of assigned BTS.) Afterwards, the addition of all cells over all types will provide the total required number of network elements in the PLMN. Finally, by the division between these numbers of elements into the total traffic managed by the network, we obtain the use factor of each network element by the traffic unit, and the unit cost can be derived.

The complete projection process is divided into two phases. Initially, the amount of BSCs is calculated; afterwards, the contributions of the rest of network elements are obtained.

BSC Projection Model

The objective of this model is to obtain the number of BSCs—that is, the use factor of the BSC—by each specific type of city. Each city considered has a heterogeneous cell deployment. This means that there is not a single type of BTS providing service, but several types distributed over the city. Using the same argument, the BTS assigned to a BSC may be of different types. The optimal case to calculate the use factor of a BSC for each city happens when all BTS of the city belongs to the same type, because it is reduced to a single division. Otherwise, we have to proceed as follows: Initially, the number of BTS is obtained under the condition that the complete city area is covered by the same type of BTS, using the following equation:

$$N_{BTS\_i} = \left\lceil \frac{City\_Area}{BTS\_i\_Coverage} \right\rceil$$

where the term $City\_Area$ is the extension of the city in Km^2. Obviously, the coverage of the BTS must be expressed in the same units.

The number of BSC to provide service to the BTS previously calculated is obtained considering several restrictions, such as the number of interfaces in the BSC, the number of active connections, the maximum traffic handled by the BSC or the link and path reliability. Afterwards, the BSC use factor for the specific type of BTS in the corresponding city is calculated as follows.

$$f\_use\_BSC_{BTS\_i} = \frac{N_{BSC}}{N_{BTS\_i}}$$

The total number of BSC in the city is calculated using the following equation:

$$N^{o}\, BSC\_City = \sum_{i=1}^{Types\_BTS} f\_use\_BSC_{BTS\_i} \cdot N_{BTS\_i}$$

MSC and NSS Projection Model

The MSC and NSS projection model is based on the same concept as the projection model of the BSC as shown in Figure 5.

The first step calculates the MSC use factor, $f\_use\_MSC_{BSC,}$ by each BSC. Afterwards, using the parameter $f\_use\_BSC_{BTS\_i}$, the use factor of the MSC for each type of considered BTS is obtained by the multiplication of both factors. Similar procedure is performed for each network element of the NSS[4].

The BSCs are connected to the MSC using optical rings usually based on STM-1[5] and STM-4 SDH systems. The number of BSCs assigned to each MSC is limited, between other factors, by the traffic capacity of the MSC and the number of interfaces towards the BSC. Therefore, the use factor of the MSC that corresponds to each BSC is calculated as follows:

$$f\_use\_MSC_{BSC} = \frac{N_{MSC}}{N_{BSC}}$$

And the MSC use factor for each type of BTS is calculated using the following equation:

$$f\_use\_MSC_{BTS\_i} = f\_use\_MSC_{BSC} \cdot f\_use\_BSC_{BTS\_i}$$

Using this methodology, an accurate estimation of the total amount of equipment for each network element on a national level can be calculated. Obviously, it is not a real configuration, but it provides a realistic structure to calculate the unit cost under the LRIC perspective.

A real example of this model application is the comparison between the investment of three different GSM operators on a limited scenario of a medium city[6]. The operators work on different bands – the first at 900 MHz, the second at 1800 MHz and the third at a double band (900 and 1800 MHz), with different types of BTSs. Hypothetical costs are assigned to each different network element under a real perspective, which means that the results can be extrapolated to practical cases. Under these premises, the differences between the operators are shown in Figure 6. The complete example is described in Fiñana (2004).

Figure 5. MSC to cell projection model scheme

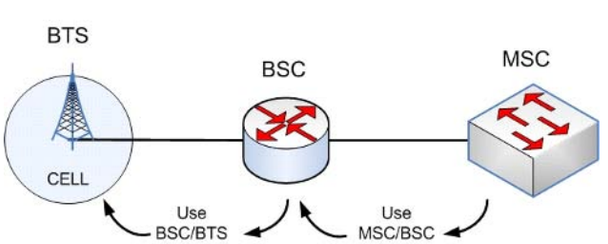


Figure 6. Comparison between the investment cost for the different operators

It can be observed that the operator with the double frequency band obtains better results in terms of investment costs. Specifically, the investment cost of this operator is 46% lower than the operator at 1800 MHz and 21% lower than the operator at 900MHz. Therefore, it has a strategic advantage that the corresponding national regulatory authority might consider on the corresponding assessment process, such as the assignment of the provision of the universal service[7] (NetworkNews, 2003).

CONCLUSION

The article has exposed a relevant problem in the current telecommunication market, which is the establishment of telecommunication services prices under a free competitive market but also under the

watchful eye of the national regulatory authorities. The two most relevant cost models have been introduced, with a deeper explanation about the LRIC model, which is currently the most widely accepted. The application of this cost model requires the complete design of the network under some restrictions, forecast of future demand, using the most suitable technology and so on. This article also deals with a possible methodology to apply the LRIC model to the GSM networks, with its particular characteristics. Finally, a short example of the relevance of this type of studies is shown, with the comparison between GSM operators working in different frequency bands.

Last, it is important to mention the relevance of this type of study, on one hand, because an erroneous network costing can establish non-realistic service prices. If they are too low, they will directly affect service profitability. If they are too high, they may reduce the number of customers and hence, affect profitability. On the other hand, under the regulation scope, these studies are required to fix an objective basis for establishing corresponding prices and, hence, to spur free competition, which is evidently the key for the telecommunication market evolution [8].

ACKNOWLEDGEMENTS

The work and results outlined in this article are performed by means of the Network of Excellence *Euro-NGI, Design and Engineering of the Next Generation Internet*, IST 50/7613 of the VI Framework of the European Community.

REFERENCES

COST 231. (1991). *Urban transmission loss models for mobile radio in the 900 and 1800MHZ bands.* Report of the EURO-COST 231 project, Revision 2.

Courcoubetis, C., & Weber, R. (2003). *Pricing communication networks.* John Wiley & Sons.

European Commission. (1998, April 8). *European Commission recommendation about interconnections. Second part: Cost accounting and account division* (Spanish). DOCE L 146 13.5.98, 6-35.

Fiñana, D., Portilla, J.A., & Hackbarth K. (2004). *2nd generation mobile network dimensioning and its application to cost models* (Spanish). University of Cantabria.

Garrese, J. (2003). The mobile commercial challenge. *Proceedings of the 42nd European Telecommunication Congress.*

Hackbarth, K., & Diallo, M. (2004). *Description of current cost and payment models for communication services and networks.* 1ST Report of Workpackage JRA 6.2 of the project Euro-NGI, IST 50/7613.

Hackbarth, K., Kulenkampff G., Gonzalez F., Rodriguez de Lope L., & Portilla, J.A. (2002). Cost and network models and their application in telecommunication regulation issues. *Proceedings of the International Telecommunication Society Congress, ITS 2002.*

Hernando-Rábanos, J.M. (1999). *Mobile communication GSM* (Spanish). Airtel Foundation.

Maciel, L.R., Bertoni, H.L., & Xia, H. (1993). Unified approach to prediction of propagation over buildings for all ranges of base station antenna height. *IEEE Transactions on Vehicular Technology, 42*(1), 41-45.

Mitchell, B., & Vogelsang, I. (1991). *Telecommunication pricing theory and practice.* Cambridge University Press.

NetworkNews. (2003, September). AMENA dominant operator in the interconnection market according to the CMT (Spanish). Redes & Telecom. Retrieved from *www.redestelecom.com*

Okumura, Y, Ohmuri E., Kawano T., & Fukuda K. (1968). Field strength and its variability in VHF and UHF land mobile service. *Review Electrical Communication Laboratory, 16*(9-10), 825-873.

Osborne, M., & Rubenstein, A. (1994). *A course on game theory*. Cambridge, MA: MIT Press.

Rappaport, S., & Hong, D. (1986). Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and non prioritized handoff procedures. *IEEE Transactions on Vehicular Technology, VT-35*(3), 77-92.

Taschdjian, M. (2001). *Pricing and tariffing of telecommunication services in Indonesia: Principles and practice.* Report of the Nathan/Checchi Joint Venture/PEG Project for the Agency for International Development.

Ward, K. (1991). Network life cycles. *Proceedings of the centennial scientific days of PKI,* Budapest.

KEY TERMS

Base Station Controller (BSC): Is the intelligent element of the Base Station Subsystem. It has complex functions in the radio resource and traffic management.

Base Station Transceiver (BTS): Is the first element that contacts the mobile terminal in the connection, and the first element of the fixed part of the mobile network.

Common Costs: It refers to the cost of joint production of a set of services.

Current Cost: It reflects the cost of the network investment over time, considering issues like amortization.

Global System for Mobile Communication (GSM): It is the second generation of mobile technology in Europe.

Historical Costs: This type of cost reflects the price of the network equipment at the time of acquisition.

Incremental Cost: It is defined as the cost of providing a specific service over a common network structure.

Mobile Switching Center (MSC): It is the switching facility of the mobile network performing the routing function using the information provided by the different database of the PLMN.

Public Land Mobile Network (PLMN): It usually means the whole network of a GSM operator.

Quality of Service (QoS): It is a mixture of several parameters, such as the ratio of server/lost calls, the quality of the service (typically voice service) in terms of noise, blur and so on. In the end, it is an objective measure of the satisfaction level of the user.

ENDNOTES

1 There are some specific exceptions to this affirmation. There is a limited set of new services in GSM, like SMS lotteries, rings and melodies downloading that are providing high revenues to the operators. Other side GPRS systems are getting some relevance, but under expectations.

2 Current costs reflect the cost of the network investment over time. Historical costs consider the equipment cost at time of acquisition.

3 The term "umbrella cells" refer to a second level of cell deployment that recovers the traffic overflowed from the cells of the initial deployment.

4 Concerning the NSS, this article limits explaining the calculation of the factor $f\_use\_MSC_{BSC}$. The use factor of the rest of elements of the NSS is calculated similarly.

5 STM-N: This term means the different transport systems of the Synchronous Digital Hierarchy network (SDH).

6 This example is performed using software named GSM-CONNECT, developed by the Telematic Engineering Group of the University of Cantabria.

7 An example is the Spanish operator AMENA, which gets only a unique frequency band at 1800 MHz. The Spanish NRA, the Comisión del Mercado de las Telecomunicaciones (CMT), has nominated it as dominant operator with the duty of providing the universal service.

8 In the field of Next Generation Internet, pricing and costing issues have a large relevance. In fact, the project IST- Euro-NGI of the VI European Framework has a specific activity oriented to this field. (see Hackbarth, 2004).

Critical Issues in Global Navigation Satellite Systems

Ina Freeman
University of Birmingham, UK

Jonathan M. Auld
NovAtel Inc., Canada

THE EVOLUTION OF GLOBAL NAVIGATION SATELLITE SYSTEMS

Global Navigation Satellite Systems (GNSS) is a concept that relays accurate information of a position or location anywhere on the globe using a minimum of four satellites, a control station, and a user receiver. GNSS owes its origins to Rabi's work in the early 1940s with the concept of an atomic clock (Nobel Museum, http://www.nobel.se/physics/laureates/1944/rabi-bio.html). In October 1940, the National Defense Research Council in the U.S. recommended implementing a new navigation system that combined radio signals with this new technology of time interval measurements. From this, MIT developed Long Range Radio Aid to Navigation (LORAN), which was refined by scientists at John Hopkins University and utilized during World War II through the late 1950s.

Following World War II, the cold war between the U.S. and the USSR embraced GNSS. The world first witnessed the emergence of the space segment of a GNSS system with the Russian Global Navigation Satellite System (GLONASS), which launched the first ICBM missile that traveled for 8,000 kilometers and the first Sputnik satellite in 1957. During this time, Dr. Ivan Getting, a man commonly noted as the father of Global Positioning System (GPS) (Anonymous, 2002), left the U.S. Department of Defense to work with Raytheon Corporation and incorporated Einstein's conceptualization of time and space into a guidance system for intercontinental missiles. Using many of these concepts, Getting worked on the first three-dimensional, time-difference-of-arrival position-finding system, creating a solid foundation for GNSS (http://www.peterson.af.mil). In 1960, Getting became the founding president of Aerospace Corp., a nonprofit corporation that works with the U.S. Department of Defense to conduct research. Getting's ongoing research resulted in a navigation system called TRANSIT, developed in the 1950s and deployed in 1960, using Doppler radar (Anonymous, 1998) and proving its effectiveness with the discovery of a Soviet missile shipment resulting in the Cuban Missile Crisis of October 1962. With the success of this system, the U.S. Secretary of Defense formed a Joint Program Office called NAVSTAR in 1973 with the intent of unifying navigation assistance within one universal system. In December 1973, the Department of Defense and the Department of Transportation published a communiqué announcing joint management of the program due to increased civilian use. Today, this is known as the Interagency GPS Executive Board (IGEB). In December 1973, the Defense System Acquisition and Review Council approved the Defense Navigation Satellite System proposal for a GNSS system, resulting in the first satellite (Navigation Technology Satellite 1—NTS1), launching on July 14, 1974 and carrying two atomic clocks (rubidium oscillators) into space. The first NAVSTAR satellites were launched in 1978 (Anonymous, 1998).

The launching of satellites in the U.S. continued until 1986, when the first *Challenger* space shuttle disaster cancelled the schedule but revived in 1989 with changes in the design of the satellite constellation, allowing enhanced access to the GNSS system by non-military users. The U.S. Coast Guard was appointed as the responsible party representing the Department of Transportation for civilian inquiries into the NAVSTAR system, resulting in the first handheld GNSS receiver, marketed by Magellan Corporation in 1989.

In January 1991, the armed conflict in Operation Desert Storm saw the GNSS system in a critical field operations role (Anonymous, 1998). Partially due to Raytheon's declaration of success in this conflict,

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the U.S. Secretary of Defense's Initial Operational Capability (IOC) recognized some of the flaws, including the inadequate satellite coverage, in the Middle East and called for improvement of the system and the resumption of research. On June 26, 1993, the U.S. Air Force launched into orbit the 24th Navstar Satellite, completing the requisites for the American GNSS system.

On July 9, 1993, the U.S. Federal Aviation Administration (FAA) approved in principle the use of the American GNSS for civil aviation. This cleared the way for the use of the system for a three-dimensional location of an object. The first official notification of this was in the February 17, 1994, FAA announcement of the increasing reliance on GNSS for civil air traffic. In 1996, a Presidential Decision Directive authorized the satellite signals to be available to civil users, and on May 2, 2000, Selective Availability was turned off, improving performance from approximately 100 meters accuracy to 10-15 meters. With the anticipated modernization of the constellation to add a third frequency to the satellites, the accuracy of the system will be enhanced to a few meters in real time. As of 2004, GPS has cost the American taxpayers $12 billion (Bellis, 2004).

THE GLOBAL GROWTH OF GNSS

Since the dissolution of the USSR, the GLONASS system has become the responsibility of the Russian Federation, and on September 24, 1993, GLONASS was placed under the auspices of the Russian Military Space Forces. The Russian government authorized civilian utilization of GLONASS in March 1995. This system declined (Langley, 1997) and did not evolve, making the system questionable for civilian or commercial use (Misra & Enge, 2001). Recognizing this, the European Union announced its intent to develop a separate civilian system known as Galileo. In 2004, the Russian government made a commitment to bring back GLONASS to a world-class system and has increased the number of functional satellites to 10 with more anticipated to a level concurrent with the American GPS.

Today, other countries of the world have recognized the importance and commercial value of GNSS and are taking steps to both broaden the technology and utilize it for their populations. The European Space Agency (ESA) has entered the second development phase to make Galileo interoperable with the U.S. GPS by developing appropriate hardware and software. Its first satellite launch is scheduled for 2008. China, India, Israel, and South Africa all have expressed an interest in joining Europe in developing the 30-satellite Galileo GNSS under the auspices of the Galileo Joint Undertaking (GJU), a management committee of the European Space Agency and the European Commission. The Japanese government is exploring the possibility of a Quazi-Zenith system that will bring the number of GNSS globally to four in the next 10 years. Thus, the globalization of navigation and positioning standards is progressing, albeit under the watchful eye of the United States, who may fear a weakening of its military prowess, and of Europe, who wants sovereign control of essential navigation services.

THE MECHANICS OF GNSS

GNSS requires access to three segments: specialized satellites in space (space segment); the operational, tracking, or control stations on the ground (control segment); and the appropriate use of localized receiver equipment (user segment). The following diagrammed system (NovAtel, Inc. Diagrams) uses a plane as the user, but it could be any user, as the same mechanics apply throughout all systems:

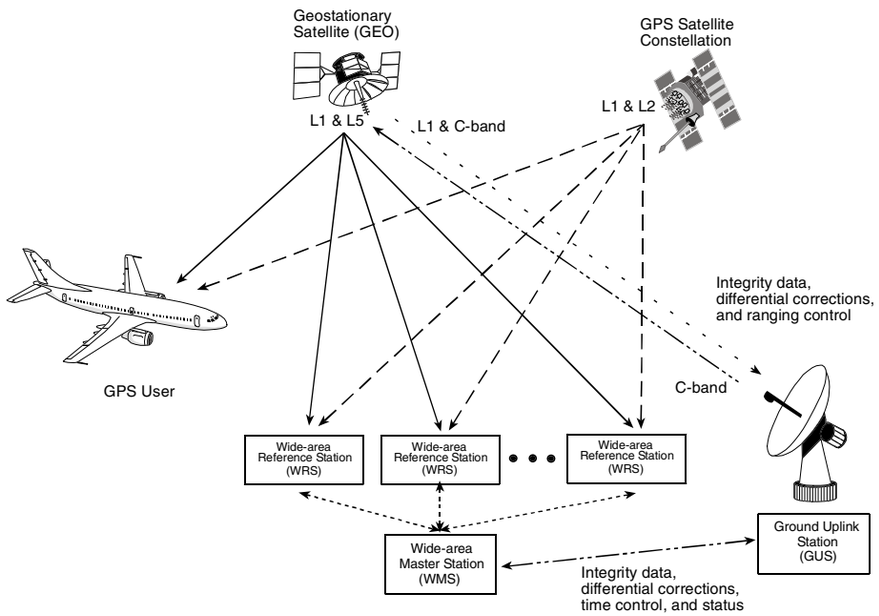
The following is a description of GPS; however, the same principles apply to all GNSSs.

Space Segment

GPS relies on 24 operational satellites (a minimum of four satellites in six orbital planes, although there are often more, depending upon maintenance schedules and projected life spans) that travel at a 54.8-degree inclination to the equator in 12-hour circular orbits, 20,200 kilometers above earth (http://www.space-technology.com/projects/gps/). They are positioned so there are usually six to eight observable at any moment and at any position on the face of the earth. Each carries atomic clocks that are accurate to within one 10-billionth of a second and broadcast signals on two frequencies (L1 and L2) (Anonymous, 1998).

The satellite emits a Pseudo Random Code (PRC)

Figure 1. GPS system



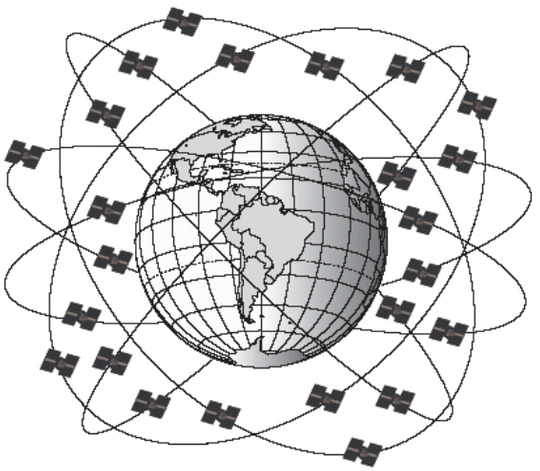
that is a series of on and off pulses in a complex pattern that reduces the likelihood of emulation or confusion of origin and that uses information theory to amplify the GPS signal, thus reducing the need for large satellite dishes. The PRC calculates the travel time of a signal from a satellite. Furthermore, each satellite broadcasts signals on two distinct frequencies that are utilized to correct for ionospheric distortions. The signals are received and identified by their unique patterns. Receivers on the earth's surface then use these signals to mathematically determine the position at a specific time.

Because the GPS signals are transmitted through two layers of the atmosphere (troposphere and ionosphere), the integrity and availability of the signal will vary (Naim, 2002). A Satellite Based Augmentation System (SBAS) augments the GPS system. This system monitors the health of the GPS satellites, provides corrections to users of the system, and is dependent upon geosynchronous satellites to provide data to the user. The SBAS system relies on the statistical principle that the more measurements taken, the greater the probability of accuracy, given consistency of all other parameters. The U.S. has structured an SBAS that is referred to as the Wide Area Augmentation System (WAAS) for use by the commercial aviation community. Work is currently underway to further enhance the WAAS with both lateral (LNAV) and vertical navigation (VNAV) capabilities, specifically useful in aviation (Nordwall, 2003). The U.S. also is investigating the capability of a Ground Based Augmentation System (GBAS) called a Local Area Augmentation Systems (LAAS) that would further enhance aviation by allowing instrument landings and take-offs under all weather conditions. With further research and reduction of costs, it could be more widely spread. SBASs are being and have been structured in various countries (e.g., the Indian GPS and GEO [Geosynchronous] Augmented Navigation (GAGAN) (anticipated); the Japanese MTSAT [Multifunction Transport Satellite] Satellite Augmentation Service [MSAS]; the Chinese Satellite Navigation Augmentation Service [SNAS]; the Canadian WAAS [CWAAS] [anticipated]; and the European Geostationary Navigation Overlay Service [EGNOS]). The satellites cover all areas of the globe, diagrammed as follows (NovAtel Inc., 2004):

Control Segment

The accuracy and currency of the satellites' functioning and orbits are monitored at a master control station operated by the 2nd Satellite Control Squadron

Figure 2. Satellite orbits



at Schriever Air Force Base (formerly Falcon), Colorado, which also operates five monitor stations dispersed equally by longitude at Ascension Island; Diego Garcia; Kwajalein, Hawaii; and Colorado Springs; and four ground antennas colocated at Ascension Island; Diego Garcia; Cape Canaveral; and Kwajalein. The incoming signals all pass through Schriever AFB where the satellite's orbits and clocks are modeled and relayed back to the satellite for transmission to the user receivers (Misra & Enge, 2001).

User Segment

The civilian use of the GPS system was made possible through the evolution of miniaturization of circuitry and by continually decreasing costs, resulting in more than 1,000,000 civilian receivers in 1997 (Misra & Enge, 2001). This technology has revolutionized the ability of the man-on-the-street to do everything from telling time and location with a wristwatch to locating a package en route.

The receiver plays a key role in the Differential GPS (DGPS) used to enhance the accuracy calculated with GPS. DGPS uses a reference station receiver at a stationary surveyed location to measure the errors in the broadcast signal and to transmit corrections to other mobile receivers. These corrections are necessitated by a number of factors, including ionosphere, troposphere, orbital errors, and clock errors. The positioning accuracy achievable can range from a few meters to a few centimeters, depending on the proximity of the receiver.

There are a number of different types of GNSS systems, all operating on the same concept but each having different levels of accuracy, as indicated in the following table.

USES OF GPS

Navigation

On June 10, 1993, the U.S. Federal Aviation Administration, in the first step taken toward using GPS instead of land-based navigation aids, approved the use of GPS for all phases of flight. Since that time, GPS has been used for flight navigation, enhancing conservation of energy. In-flight navigation was integrated with take-offs and landings in September 1998, when the Continental Airlines Airbus MD80 used GPS for all phases of flight (Bogler, 1998). This technology also allows for more aircraft to fly the skies, because separation of flight paths is possible with the specific delineation capable only with GPS. Further, GPS is integral to sea navigation for giant ocean-going vessels in otherwise inaccessible straights and through passages to locate prime fishing areas and for small fishing boats.

Navigation is not restricted to commercial enterprises. Individuals such as hikers, bikers, mountaineers, drivers, and any other person who may traverse distance that is not clearly signed can use GPS to find their way.

Survey/Location

GNSS can be used to determine the position of any vehicle or position. Accuracy can be as high as 1-2 centimeters with access to a reference receiver and transmitter. GPS is currently an integral part of search and rescue missions and surveying (in particular, it is being used by the Italian government to create the first national location survey linked to the WGS-84 coordinated reference frame). Receivers are used by geologists, geodesists, and geophysicists to determine the risk of earthquakes, sheet ice movement, plate motion, volcanic activity, and variations in the earth's

Table 1. GNSS system

Type of System	Positioning Type	Accuracy	Coverage
GPS	Stand Alone – no external augmentation	5 meters	Global
DGPS	Differential GPS	~1 – 2 meters	Typically less than 100 km relative to correction source
RTK	Precise Differential GPS	~ 1 – 2 cm	Typically less than 40 km relative to correction source
SBAS	GPS augmented by network correction from GEO Stationery Satellite	~ 1 – 2 meters	National and/or continental
E-911 (E-OTD)	Cellular Network Based Positioning	50 – 150 meters	Dependent on size of network – typically mobile phone dependent
AGPS	Assisted GPS – based on GPS but augmented by data from cellular network	50 – 150 meters	Dependent on size of network – typically mobile phone dependent
LORAN	Ground/Land Based Navigation	450 meters	Typically national – dependent on network size

rotation; by archeologists to locate and identify difficult to locate dig sites; by earth movers to determine where to work; and by farmers to determine their land boundaries. In the future, self-guided cars may become a reality.

Asset Tracking

The world of commerce uses GPS to track its vehicle fleets, deliveries, and transportation scheduling. This also includes tracking of oil spills and weather-related disasters such as flooding or hurricanes, and tracking and locating containers in storage areas at ports.

E-911

This system began with the primary purpose of locating people who contacted emergency services via cell phones. It now also tracks emergency vehicles carrying E-OTD (Enhanced Observed Time Difference) equipment to facilitate the determination of the fastest route to a specified destination.

Mapping

GPS can be used as easily to explore locations as it can be used to locate people and vehicles. This ability enhances the accuracy of maps and models of everything from road maps to ecological surveys of at-risk animal populations to tracking mountain streams and evaluating water resources both on earth and in the troposphere.

Communication

GPS can be coordinated with communication infrastructures in many applications in transportation, public safety, agriculture, and any primary applications needing accurate timing and location abilities. GPS facilitates computerized and human two-way communication systems such as emergency roadside assistance and service, enhancing the speed of any transaction.

Agriculture

GPS is used in agriculture for accurate and efficient application of fertilizer or herbicides and for the mechanized harvesting of crops.

Construction

With the use of DGPS, construction may be completed using CAD drawings without manual measurements, reducing time and costs. Currently, GPS

assists in applications such as setting the angle of the blade when digging or building a road. It also assists with monitoring structural weaknesses and failures in engineering structures such as bridges, buildings, towers, and large buildings.

Time

With the use of two cesium and two rubidium atomic clocks in each satellite and with the automatic corrections that are part of the system, GPS is an ideal source of accurate time. Time is a vital element for many in commerce, including banks, telecommunication networks, power system control, and laboratories, among others. It is also vital within the sciences, including astronomy and research.

Miscellaneous

The uses of GPS are varied and include individually designed applications such as the tracking of convicts in the community via the use of an ankle band, the location of high-value items such as the bicycles used in the Tour de France, and surveillance.

CONCLUSION: WHERE TO FROM HERE

The future of GNSS is emerging at a phenomenal pace. Already in prototype is a new GPS navigation signal at L5. When used with both WAAS and LAAS, this will reduce ionospheric errors and increase the integrity of the received data. The introduction of the interoperable Galileo system will enhance further the GPS system and further refine the precision of the measurements. Commerce continually speaks of globalization and the positive effects of this phenomenon upon humankind. With the increasing usage of GNSS systems, this globalization becomes a seamless system with governments and private enterprises interacting across national borders for the benefit of all. As commercial enterprises around the world become increasingly dependent on GNSS, these invisible waves may bring together what governments cannot.

REFERENCES

Anonymous. (1998). Global positioning system marks 20th anniversary. *Program Manager, 27*(3), 38-39.

Anonymous. (2002). Dr. Ivan A. Getting genius behind GPS. *Business & Commercial Aviation, 91*(6), 76.

Bellis, M. (2004). Inventors: Global positioning system—GPS. Retrieved August 2004, from *http://inventors.about.com/library/inventors/blgps.htm*

Bogler, D. (1998). Precision landing ready for take off: Technology aviation: Aircraft congestion may soon be a thing of the past. *Financial Times,* 18.

GPS World. *http://www.gpsworld.com/gpsworld/static/staticHtml.jsp?id=2294*

Hasik, J., & Rip, M. (2003). An evaluation of the military benefits of the Galileo system. *GPS World, 14*(4), 28-31.

Interagency Executive GPS Executive Board. *http://www.igeb.gov/*

Langley, R.B. (1997). GLONASS: Review and update. *GPS World, 8*(7), 46-51.

Misra, P., & Enge, P. (2001). *Global positioning system: Signals, measurements, and performance.* Lincoln, MA: Ganga-Jamuna Press.

Naim, G. (2002). Technology that watches over us: Satellite-based air traffic control. *Financial Times,* 3.

Nobel Museum. (n.d.). *Isadore Isaac Rabi Biography*. Retrieved August 2004, from *http://www.nobel.se/physics/laureates/1944/rabi-bio.html*

Nordwall, B.D. (2003). GNSS at a crossroads capstone shows what WAAS can do in Alaska. Are LAAS and Galileo far behind? And what will other global nav satellite systems bring? *Aviation Week & Space Technology, 15*9(10), 58.

NovAtel, Inc. (2004). Documents and Graphics. Permission received from CEO Mr. Jonathan Ladd.

U.S. Coast Guard Navigation Center. *http://www.navcen.uscg.gov/gps/default.htm*

KEY TERMS

Ephemeris Data Parameters: Ephemeris data parameters describe short sections of the space vehicle or satellite orbits. New data are gathered by receivers each hour. However, the receiver is capable of using data gathered four hours before without significant error. Algorithms are used in conjunction with the ephemeris parameters to compute the SV position for any time within the period of the orbit described by the ephemeris parameter set.

Ionosphere: A band of particles 80-120 miles above the earth's surface.

LAAS: Local Area Augmentation System. A safety-critical navigation system that provides positioning information within a limited geographic area.

Pseudo Random Noise (PRN): PRN is a noise-like series of bits. Because GNSS depends upon multiple inputs, each satellite produces a predetermined, unique PRN on both the L1 and the L2 carrier signal for use by civil and military receivers. The L2 carrier signal is restricted to military use.

Standard Positioning Service (SPS): The signal that is available to civil users worldwide without charge or restrictions and is available/usable with most receivers. The U.S. DoD is responsible for the transmission of this data. U.S. government agencies have access to Precise Positioning Service (PPS), which uses equipment with cryptographic equipment and keys and specially equipped receivers that have the capability of using the exclusive L2 frequency for enhanced information. The PPS gives the most accurate dynamic positioning possible.

Systems: Systems are being deployed by various political entities to form a global network. GLONASS (Global Navigation Satellite System) is deployed by the Russian Federation. GPS **(**Global Positioning System) is deployed by the United States. Galileo is the GPS system being structured by the European Union.

Troposphere: The densest part of the earth's atmosphere that extends from the surface of the earth to the bottom of the stratosphere and in which most weather changes occur and temperature fluctuates.

WAAS: Wide Area Augmentation System. A safety-critical navigation system that provides positioning information.

Dark Optical Fibre as a Modern Solution for Broadband Networked Cities

Ioannis P. Chochliouros
Hellenic Telecommunications Organization S.A. (OTE), Greece

Anastasia S. Spiliopoulou-Chochliourou
Hellenic Telecommunications Organization S.A. (OTE), Greece

George K. Lalopoulos
Hellenic Telecommunications Organization S.A. (OTE), Greece

INTRODUCTION: BROADBAND PERSPECTIVE

The world economy is moving in transition from the industrial age to a new set of rules—that of the so-called information society—which is rapidly taking shape in different multiple aspects of everyday life: The exponential growth of the Internet, the explosion of mobile communications, the rapid emergence of electronic commerce, the restructuring of various forms of businesses in all sectors of the modern economy, the contribution of digital industries to growth and employment, and so forth are some amongst the current features of the new global reality. Changes are usually underpinned by technological progress and globalisation, while the combination of global competition and digital technologies is having a crucial sweeping effect. Digital technologies facilitate the transmission and storing of information while providing multiple access facilities, in most cases, without significant costs. As digital information may be easily transformed into economic and social value, it offers huge opportunities for the development of new products, services, and applications. Information becomes the key resource and the engine of the new e-economy. Companies in different sectors have started to adapt to the new economic situation in order to become e-businesses (European Commission, 2001c). The full competitiveness of a state in the current high-tech, digitally converging environment is strongly related to the existence of a modern digital infrastructure of high capacity and high performance that is rationally deployed, properly priced, and capable of providing easy, cost-effective, secure, and uninterrupted access to the international "digital web" of knowledge and commerce without imposing any artificial barriers and/or restrictions.

Broadband development is a major priority for the European Union (EU) (Chochliouros & Spiliopoulou-Chochliourou, 2003a). Although there is still a crisis in the sector, the information society is still viewed as a powerful source of business potential and improvements in living standards (European Commission, 2001b). To appropriate further productivity gains, it should be necessary to exploit the advances offered by the relevant technologies, including high-speed connections and multiple Internet uses (European Commission, 2002). To obtain such benefits, it should be necessary to develop new cooperative and complementary network facilities. Among the various alternatives, Optical-Access Networks (OANs) can be considered, for a variety of explicit reasons, as a very reliable solution, especially in urban areas.

The development of innovative communications technologies, the digital convergence of media and content, the exploitation and penetration of the Internet, and the emergence of the digital economy are main drivers of the networked society, while significant economic activities are organized in networks (including development and upgrading), especially within urban cities (European Commission, 2003). In fact, cities remain the first interface for citizens and enterprises with the administrators and the main providers of public services. In recent years, there have been significant advances in the speed and the capacity of Internet-based backbone networks, including those of fibre. Furthermore, there is a strong challenge for the exploitation of

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

dark fibre infrastructure and for realising various access networks. Such networks are able to offer an increase in bandwidth and quality of service for new and innovative multimedia applications.

NETWORKED CITIES: TOWARD A GLOBAL AND SUSTAINABLE INFORMATION SOCIETY

Information society applications radically transform the entire image of the modern era. In particular, a great variety of innovative electronic communications and applications provide enormous facilities both to residential and corporate users (European Commission, 2001a), while cities and regions represent major "structural" modules. Local authorities are key players in the new reality as they are the first level of contact between the citizens and the public administrations and/or services. Simultaneously, because of the new information geography and global economy trends, they act as major "nodes" in a set of interrelated networks where new economic processes, investments, and knowledge take place. Recently, there is a strong interest for cooperation between global and local players (through schemes of private or public partnerships) in major cities of the world, especially for the spread of knowledge and technology. Encouraging investment in infrastructure (by incumbent operators and new entrants) and promoting innovation are basic objectives for further development.

In particular, the deployment of dark-fibre-optics infrastructure (Arnaud, 2000) under the form of Metropolitan Area Networks (MANs) can guarantee an effective facilities-based competition with a series of benefits. It also implicates that, apart from network deployment, there would be more and extended relevant activities realised by other players, such as Internet Service Providers (ISPs), Application Service Providers (ASPs), operators of data centres, and so forth. Within the same framework, of particular importance are business opportunities, especially for the creation of dark customer-owned infrastructure and carrier "neutral" collocation facilities.

In recent years, there have been significant advances in the speed and capacity of Internet backbone networks, including those of fibre-based infrastructure. These networks can offer an increase in bandwidth and quality of service for advanced applications. At the same time, such networks may contribute to significant reductions in prices with the development of new (and competitive) service offerings. In the context of broadband, local decision-making is extremely important. Knowledge of local conditions and local demand can encourage the coordination of infrastructure deployment, providing ways of sharing facilities (European Parliament & European Council, 2002a) and reducing costs. The EU has already proposed suitable policies (Chochliouros & Spiliopoulou-Chochliourou, 2003d) and has organized the exchange of best practices at the total, regional, and local level, while expecting to promote the use of public and private partnerships.

At the initial deployment of fibre in backbone networks, there was an estimate that fibre could be deployed to the home as well. A number of various alternate FTT*x* schemes or architecture models such as fibre to the curb (FTTC), fibre to the building (FTTB), fibre to the home (FTTH), hybrid fibre coaxial (HFC), and switched digital video (SDV) have been introduced (Arnaud, 2001) and tested to promote not only basic telephony and video-on-demand (VOD) services, but broadband applications as well. Such initiatives have been widely developed by telecommunications network operators.

DARK FIBRE SOLUTIONS: CHALLENGES AND LIMITATIONS

Apart from the above "traditional" fibre-optic networks, there is a recent interest in the deployment of a new category of optical networks. This originates from the fact that for their construction and for their effective deployment and final use, the parties involved generate and promote new business models completely different from all the existing ones. Such models are currently deployed in many areas of North America (Arnaud, 2002). As for the European countries, apart from a successful pilot attempt in Sweden (STOKAB AB, 2004), such an initiative is still "immature". However, due to broadband and competition challenges, such networks may provide

valuable alternatives for the wider development of potential information society applications, in particular, under the framework of the recent common EU initiatives (Chochliouros & Spiliopoulou-Chochliourou, 2003a; European Commission, 2001b; European Commission, 2002).

"Dark fibre" is usually an optical fibre dedicated to a single customer, where the customer is responsible for attaching the telecommunications equipment and lasers to "light" the fibre (Foxley, 2002). In other words, a "dark fibre" is an optical fibre without communications equipment; that is, the network owner gives both ends of the connection in the form of fibre connections to the operator without intermediate equipment. In general, dark fibre can be more reliable than traditional telecommunications services, particularly if the customer deploys a diverse or redundant dark fibre route. This option, under certain circumstances, may provide incentive for further market exploitation and/or deployment while reinforcing competition (Chochliouros & Spiliopoulou-Chochliourou, 2003c). Traditionally, optical-fibre networks have been built by network operators (or "carriers") who take on the responsibility of lighting the relevant fibre and provide a managed service to the customer.

Dark fibre can be estimated, explicitly, as a very simple form of technology, and it is often referred to as technologically "neutral". Sections of dark fibre can be very easily fused together so that one continuous strand exists between the customer and the ultimate destination. As such, the great advantage of dark fibre is that no active devices are required in the fibre path. Due to the non-existence of such devices, a dark fibre in many cases can be much more reliable than a traditional managed service. Services of the latter category usually implicate a significant number of particular devices in the network path (e.g., ATM [Asynchronous Transfer Mode] switches, routers, multiplexers, etc.); each one of these intermediates is susceptible to failure and this becomes the reason why traditional network operators have to deploy complex infrastructure and other systems to assure compatibility and reliability. For the greatest efficiency, many customers usually select to install two separate dark fibre links to two separate service providers; however, even with an additional fibre, dark fibre networks are cheaper than managed services from a network operator.

With customer-owned dark fibre networks, the end customer becomes an "active entity" who finally owns and controls the relevant network infrastructure (Arnaud, Wu, & Kalali, 2003); that is, the customers decide to which service provider they wish to connect with for different services such as, for example, telephony, cable TV, and Internet (New Paradigm Resources Group, Inc., 2002). In fact, for the time being, most of the existing customer-owned dark fibre deployments are used for delivery of services and/or applications based on the Internet (Crandall & Jackson, 2001). The dark fibre industry is still evolving.

With the dark fibre option, customers may have further choices in terms of both reliability and redundancy. That is, they can have a single unprotected fibre link and have the same reliability as their current connection (Arnaud et al., 2003; New Paradigm Resources Group, Inc., 2002); they can use alternative technology, such as a wireless link for backup in case of a fibre break; or they can install a second geographically diverse dark fibre link whose total cost is still cheaper than a managed service as indicated above. Furthermore, as fibre has greater tensile strength than copper (or even steel), it is less susceptible to breaks from wind or snow loads. Network cost and complexity can be significantly reduced in a number of ways. As already noticed, dark fibre has no active devices in the path, so there are fewer devices to be managed and less statistical probability of the appearance of fault events. Dark fibre allows an organization to centralize servers and/or outsource many different functions such as Web hosting, server management, and so forth; this reduces, for example, the associated management costs. Additionally, repair and maintenance of the fibre is usually organized and scheduled in advance to avoid the burden of additional costs. More specifically, dark fibre allows some categories of users such as large enterprise customers, universities, and schools to essentially extend their in-house Local Area Networks (LANs) across the wide area. As there is no effective cost to bandwidth, with dark fibre the long-distance LAN can still be run at native speeds with no performance degradation to the end user. This provides an option to relocate, very simply, a server to a distant location where previously it

required close proximity because of LAN performance issues (Bjerring & Arnaud, 2002).

Although dark fibre provides major incentive to challenge the forthcoming broadband evolution, it is not yet fully suitable for all separate business cases. The basic limitation, first of all, is due to the nature of the fibre, which is normally placed at "fixed" and predetermined locations. This implicates that relevant investments should be done to forecast long-term business activities. Such a perspective is not quite advantageous for companies leasing or renting office space, or that desire mobility; however, this could be ideal for organizations acting as fixed institutions at specific and predefined premises (e.g., universities, schools, hospitals, public-sector institutions, libraries, or large businesses). Furthermore, the process to deploy a dark fibre network is usually a hard task that requires the consumption of time and the resolution of a variety of problems, including technical, financial, regulatory, business, and other difficulties or limitations. Detailed engineering studies have to be completed, and municipal-access and related support-structure agreements have to be negotiated before the actual installation of the fibre begins (Chochliouros & Spiliopoulou-Chochliourou, 2003a, 2003c, 2003d; European Commission, 2001b, 2002).

Around the world, a revolution is taking place in some particular cases of high-speed networking. Among other factors, this kind of activities is driven by the availability of low-cost fibre-optic cabling. In turn, lower prices for fibre are leading to a shift from a telecommunications network operators infrastructure (or "carrier-owned" infrastructure) toward a more "customer-owned" or municipally-owned fibre, as well as to market-driven innovative sharing arrangements such as those guided by the "condominium" fibre networks. This implicates a very strong challenge, especially under the scope of the new regulatory measures (European Parliament & European Council, 2002b) for the promotion of the deployment of modern electronic communications networks and services. It should be expected that both the state (also including all responsible regulatory authorities) and the market itself would find appropriate ways to cooperate (Chochliouros & Spiliopoulou-Chochliourou, 2003b; European Parliament & European Council, 2002a) in order to provide immediate solutions.

A "condominium" fibre is a unit of dark fibre (Arnaud, 2000, 2002) installed by a particular contractor (originating either from the private or the public sector) on behalf of a consortium of customers, with the customers to be owners of the individual fibre strands. Each customer-owner lights the fibres using his or her own technology, thereby deploying a private network to wherever the fibre reaches (i.e., to any possible terminating location or endpoint, perhaps including telecommunications network operators and Internet providers). The business arrangement is comparable to a condominium apartment building, where common expenses such as management and maintenance fees are the joint responsibility of all the owners of the individual fibres. A "municipal" fibre network is a network of a specific nature and architecture (Arnaud, 2002) owned by a municipality (or a community). Its basic feature is that it has been installed as a kind of public infrastructure with the intention of leasing it to any potential users (under certain well-defined conditions and terms). Again, "lighting" the fibre to deploy private network connections is the responsibility of the lessee, not the municipality. Condominium or municipal fibre networks, due to the relevant costs as well as to the enormous potential they implicate for innovative applications, may be of significant interest for a set of organizations such as libraries, universities, schools, hospitals, banks, and the like that have many sites distributed over a distinct geographic region. The development of dark fibre networks may have a radical effect on the traditional telecommunications business model (in particular, when combined with long-haul networks based on customer-owned wavelengths on Dense-Wavelength Division-Multiplexed [DWDM] systems; Arnaud, 2000; Chochliouros & Spiliopoulou-Chochliourou, 2003d; European Commission, 2001b, 2002, 2003; European Parliament & European Council, 2002b). Such a kind of infrastructure may encourage the further spreading of innovative applications such as e-government, e-education, and e-health.

CONCLUSION

Dark fibre provides certain initiatives for increased competition and for the benefits of different categories of market players (network operators, service

providers, users-consumers, various authorities, etc.); this raises the playing field among all parties involved for the delivery of relevant services and applications. Dark fibre may strongly enable new business activities while providing major options for low cost, simplicity, and efficiency under suitable terms and/or conditions for deployment (Chochliouros & Spiliopoulou-Chochliourou, 2003c).

The dark fibre industry is still immature at a global level. However, there is a continuous evolution and remarkable motivation to install, sell, or lease such a network infrastructure, especially for emerging broadband purposes. The perspective becomes more important via the specific option of customer-owned dark fibre networks, where the end customer becomes an "active entity" who finally owns and controls the relevant network infrastructure; that is, the customers decide to which service provider they wish to connect with at a certain access point for different services such as, for example, telephony, cable TV, and Internet (Chochliouros & Spiliopoulou-Chochliourou, 2003d).

Dark fibre may be regarded as "raw material" in the operator's product range and imposes no limits on the services that may be offered. However, due to broadband and competition challenges, such networks may provide valuable alternatives for the wider development of potential information society applications. In particular, under the framework of the recent common EU initiatives for an e-Europe 2005 (European Commission, 2002), such attempts may contribute to the effective deployment of various benefits originating from the different information society technology sectors. Moreover, these fibre-based networks raise the playing field and provide multiple opportunities for cooperation and business investments among all existing market players in a global electronic communications society.

REFERENCES

Arnaud, B. S. (2000). *A new vision for broadband community networks.* CANARIE, Inc. Retrieved August 10, 2004 from *http://www.canarie.ca/canet4/library/customer.html*

Arnaud, B. S. (2001). *Telecom issues and their impact on FTTx architectural designs (FTTH Council).* CANARIE, Inc. Retrieved August 2, 2004 from *http://www.canarie.ca/canet4/library/customer.html*

Arnaud, B. S. (2002). *Frequently asked questions (FAQ) about community dark fiber networks.* CANARIE, Inc. Retrieved August 5, 2004 from *http://www.canarie.ca/canet4/library/customer.html*

Arnaud, B. S., Wu, J., & Kalali, B. (2003). *Customer controlled and managed optical networks.* CANARIE, Inc. Retrieved July 20, 2004 from *http://www.canarie.ca/canet4/library/canet4design.html*

Bjerring, A. K., & Arnaud, B. (2002). *Optical Internets and their role in future telecommunications systems.* CANARIE, Inc. Retrieved July 20, 2004 from *http://www.canarie.ca/canet4/library/general.html*

Chochliouros, I., & Spiliopoulou-Chochliourou, A. (2003a). The challenge from the development of innovative broadband access services and infrastructures. *Proceedings of EURESCOM SUMMIT 2003: Evolution of Broadband Services-Satisfying User and Market Needs* (pp. 221-229). Heidelberg, Germany: EURESCOM & VDE Publishers.

Chochliouros, I., & Spiliopoulou-Chochliourou, A. (2003b). Innovative horizons for Europe: The new European telecom framework for the development of modern electronic networks & services. *The Journal of the Communications Network: TCN, 2*(4), 53-62.

Chochliouros, I., & Spiliopoulou-Chochliourou, A. (2003c). New model approaches for the deployment of optical access to face the broadband challenge. *Proceedings of the Seventh IFIP Working Conference on Optical Network Design & Modelling: ONDM2003* (pp. 1015-1034). Budapest, Hungary: Hungarian National Council for Information Technology (NHIT), Hungarian Telecommunications Co. (MATAV PKI) and Ericsson Hungary.

Chochliouros, I., & Spiliopoulou-Chochliourou, A. (2003d). Perspectives for achieving competition and development in the European information and communications technologies (ICT) markets. *The Jour-*

nal of the Communications Network: TCN, 2(3), 42-50.

Crandall, R. W., & Jackson, C. L. (2001). The $500 billion opportunity: The potential economic benefit of widespread diffusion of broadband Internet access. In A. L. Shampine (Ed.), *Down to the wire: Studies in the diffusion and regulation of telecommunications technologies.* Haupaugge, NY: Nova Science Press. Retrieved July 15, 2004 from *http://www.criterioneconomics.com/pubs/articles_crandall.php*

European Commission. (2001a). *Communication on helping SMEs to go digital [COM (2001) 136, 13.03.2001].* Brussels, Belgium: European Commission.

European Commission. (2001b). *Communication on impacts and priorities [COM (2001) 140, 13.03.2001].* Brussels, Belgium: European Commission.

European Commission. (2001c). *Communication on the impact of the e-economy on European enterprises: Economic analysis and policy implications [COM(2001) 711, 29.11.2001].* Brussels, Belgium: European Commission.

European Commission. (2002). *Communication on eEurope 2005: An information society for all-An action plan [COM (2002) 263, 28.05.2002].* Brussels, Belgium: European Commission.

European Commission. (2003). *Communication on electronic communications: The road to the knowledge economy [COM (2003) 65, 11.02.2003].* Brussels, Belgium: European Commission.

European Parliament & European Council. (2002a). *Directive 2002/19/EC on access to, and interconnection of, electronic communications networks and associated facilities (Access directive) [OJ L108, 24.02.2002, 7-20].* Brussels, Belgium: European Commission.

European Parliament & European Council. (2002b). *Directive 2002/21/EC on a common regulatory framework for electronic communications networks and services (Framework directive) [OJ L108, 24.04.2002, 33-50].* Brussels, Belgium: European Commission.

Foxley, D. (2002, January 31). *Dark fiber.* TechTarget.com, Inc. Retrieved August 1, 2004 from *http://searchnetworking.techtarget.com//sDefinition/0,,sid7_gci21189,00.html*

New Paradigm Resources Group, Inc. (2002, February). Dark fiber: Means to a network. *Competitive Telecom Issues, 10*(2). Retrieved June 11, 2004 from *http://www.nprg.com*

STOKAB AB. (2004). *Laying the foundation for IT: Annual report 2003.* City of Stockholm, Sweden. Retrieved June 6, 2004 from *http://www.stokab.se*

KEY TERMS

Broadband: A service or connection allowing a considerable amount of information to be conveyed, such as video. It is generally defined as a bandwidth of more than 2 Mbit/s.

Carrier "Neutral" Collocation Facilities: Facilities, especially in cities, built by companies to allow the interconnection of networks between competing service providers and for the hosting of Web servers, storage devices, and so forth. They are rapidly becoming the "obvious" location for terminating "customer-owned" dark fibre. (These facilities, also called "carrier-neutral hotels", feature diesel-power backup systems and the most stringent security systems. Such facilities are open to carriers, Web-hosting firms and application service firms, Internet service providers, and so forth. Most of them feature a "meet-me" room where fibre cables can be cross-connected to any service provider within the building. With a simple change in the optical patch panel in the collocation facility, the customer can quickly and easily change service providers on very short notice.)

Condominium Fibre: A unit of dark fibre installed by a particular contractor (originating either from the private or the public sector) on behalf of a consortium of customers, with the customers to be owners of the individual fibre strands. Each customer-owner lights the fibres using his or her own technology, thereby deploying a private network to wherever the fibre reaches, that is, to any possible terminating location or endpoint.

Dark Fibre: Optical fibre for infrastructure (cabling and repeaters) that is currently in place but is not being used. Optical fibre conveys information in the form of light pulses, so "dark" means no light pulses are being sent.

Dense-Wavelength Division Multiplexing (DWDM): The operation of a passive optical component (multiplexer) that separates (and/or combines) two or more signals at different wavelengths from one (two) or more inputs into two (one) or more outputs.

FTT*x*: Fibre to the cabinet (Cab), curb (C), building (B), or home (H).

Local Area Network (LAN): A data communications system that (a) lies within a limited spatial area, (b) has a specific user group, (c) has a specific topology, and (d) is not a public-switched telecommunications network, but may be connected to one.

Metropolitan Area Network (MAN): A data network intended to serve an area approximating that of a large city. Such networks are being implemented by innovative techniques, such as running fibre cables through subway tunnels.

Municipal Fibre Network: A network of a specific nature and architecture owned by a municipality (or a community). Its basic feature is that it has been installed as a kind of public infrastructure with the intention of leasing it to any potential users (under certain well-defined conditions and terms). Again, lighting the fibre to deploy private network connections is the responsibility of the lessee, not the municipality.

Optical-Access Network (OAN): The set of access links sharing the same network-side interfaces and supported by optical-access transmission systems.

The Decision Making Process of Integrating Wireless Technology into Organizations

Assion Lawson-Body
University of North Dakota, USA

Glenda Rotvold
University of North Dakota, USA

Justin Rotvold
Techwise Solutions, LLC, USA

INTRODUCTION

With the advancement of wireless technology and widespread use of mobile devices, many innovative mobile applications are emerging (Tarasewich & Warkentin, 2002; Varshney & Vetter, 2002; Zhang, 2003). Wireless technology refers to the hardware and software that allow transmission of information between devices without using physical connections (Zhang, 2003). Understanding the different technologies that are available, their limitations, and uses can benefit companies looking at this technology as a viable option to improve overall organizational effectiveness and efficiency.

A significant part of the growth in electronic business is likely to originate from the increasing numbers of mobile computing devices (Agrawal, Kaushal, & Ravi, 2003; Anderson & Schwager, 2004; Varshney & Vetter, 2000). Ciriello (as cited in Smith, Kulatilaka, & Venkatramen, 2002, p. 468) states that "Forecasts suggest that the number of worldwide mobile connections (voice and data) will grow from 727 million in 2001 to 1,765 million in 2005." With the huge growth anticipated in the utilization of wireless technologies, businesses are going to be increasingly faced with decisions on what wireless technologies to implement.

The objective of this article is to examine and discuss wireless technologies followed by presentation and discussion of a decision model that was formed to be used in determining the appropriate wireless technology. Technologies appropriate for both mobile and wide area coverage are discussed followed by technologies such as WLANs, which are used in more local, confined areas with short to medium range communication needs.

This article is organized as follows. The first section contains the various generations of Wireless Technology; in the second, WLANs are examined. The following section describes a decision model. In the next section, technology concerns are discussed, and the final section presents the conclusion.

WIRELESS TECHNOLOGY: GENERATIONS

There has been an industry-wide understanding of different "generations" regarding mobile technology (Varshney & Jain, 2001). Currently, there are also several technologies within each classification of generations, but the technologies are not necessarily finite in these generations.

First Generation

First generation (1G) contains analog cellular systems and does not have the capability to provide data services. The only service is voice service that can be provided to mobile phones. Two technologies worth noting are advance mobile phone service (AMPS) and frequency division multiple access (FDMA). AMPS is a first generation analog cellular phone system standard that operates in the 800 Mhz band. AMPS uses FDMA (an access/multiplexing technology) which separates the spectrum into 30 kHz channels, each of which can carry a voice conversation or, with digital service, carry digital data. FDMA allows for

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

multiple users to "access a group of radio frequency bands" and helps eliminate "interference of message traffic" (Dunne, 2002).

Second Generation

Second generation (2G) is a digital wireless telephone technology that uses circuit-switched services. This means that a person using a second generation-enabled device must dial in to gain access to data communications. "Circuit-switched connections can be slow and unreliable compared with packet-switched networks, but for now circuit-switched networks are the primary method of Internet and network access for wireless users in the United States" (Dunne, 2002). In this generation one will find Global System for Mobile communications (GSM) which is a network standard, in addition to time division multiple access (TDMA) and code division multiple access (CDMA), which are multiplexing technologies. The 2G technology that is most widely used is GSM (a standard with the highest use in Europe) with a data rate of 9.6 kilobits per second (Tarasewich, Nickerson & Warkentin, 2002). TDMA works with GSM while CDMA does not, but CDMA is more widely used in the United States (Dunne, 2002).

TDMA allows many users to use the same radio frequency by breaking the data into fragments, which are each assigned a time slot (Dunne, 2002). Since each user of the channel takes turns transmitting and receiving, only one person is actually using the channel at any given moment and only uses it for short bursts. CDMA on the other hand, uses a special type of digital modulation called Spread Spectrum, which spreads the user's voice stream bits across a very wide channel and separates subscriber calls from one another by code instead of time (Agrawal et al., 2003). CDMA is used in the U.S. by carriers such as Sprint and Verizon (Dunne, 2002).

Two and One-Half Generation

There is a half generation that follows 2G. 2.5G exhibits likenesses of both 2G and 3G technologies. 2G wireless uses circuit switched connections while 3G uses high-speed packet switched transmission. Circuit-switching requires a dedicated, point to point physical circuit between two hosts where the bandwidth is reserved and the path is maintained for the entire session. Packet switching, however, divides digitized messages into packets, which contain enough address information to route them to their network destination. The circuit is maintained only as long as it takes to send the packet resulting in cost savings.

High-speed circuit-switched data (HSCSD), enhanced data GSM environment (EDGE), and general packet radio service (GPRS) exist in this generation. HSCSD is circuit switched, but can provide faster data rates of up to 38.4 Kbps, which sets it apart from 2G. EDGE separates itself from 2G by being a version of GSM that is faster and is designed to be able to handle data rates up to 384 Kbps (Tarasewich et al., 2002). GPRS uses packet switching. GPRS, a service designed for digital cellular networks, utilizes a packet radio principle and can be used for carrying end users' packet data protocol such as IP information to and from GPRS terminals and/or external packet data networks. GPRS is different by being a packet data service. A packet data service provides an "always-on" feature so users of the technology do not have to dial in to gain Internet access (Tarasewich et al., 2002). Although this technology is packet based, it still is designed to work with GSM (Dunne, 2002).

Third Generation

This generation is what will occur next. Although 3G has recently been deployed in a few locations, it is now in the process of being deployed in additional regions. This process of installation and migration to 3G will take time to completely implement on a widespread basis across all areas of the globe. There will be high-speed connections and increasing reliability in this generation that will allow for broadband for text, voice, and even video and multimedia. It utilizes packet-based transmissions as well giving the ability to be "always-on." 3G is capable of network convergence, a newer term used to describe "the integration of several media applications (data, voice, video, and images) onto a common packet-based platform provided by the Internet Protocol (IP)" (Byun & Chatterjee, 2002, p. 421). Whether or not the protocol used for packet-based transfer (on a handheld or smart phone) is the Internet Protocol, depends on the devices.

A derivative of CDMA, a wideband CDMA is expected to be developed that will require more bandwidth than CDMA because it will utilize multiple

wireless signals, but in turn, using multiple wireless signals will provide greater bandwidth (Dunne, 2002). For example, Ericsson and Japan Telecom successfully completed the world's first field trial of voice-over-IP using wideband CDMA. A technology hopeful in 3G is universal mobile telecommunications system. This is said to be the planned global standard that will provide data rates of up to and exceeding 2 Mbps (Tarasewich et al., 2002).

WIRELESS LOCAL AREA NETWORKS (WLAN)

We will now shift our focus from long-range mobile communications to technologies appropriate for short to medium range coverage areas. In fact, WLAN represents a category of wireless networks that are typically administered by organizations (Agrawal et al., 2003) and many of the issues with wireless telecommunications technologies are similar to those found with wireless LANs (Tarasewich et al., 2002).

Wireless Physical Transport

The Institute of Electrical and Electronics Engineers (IEEE) has developed a set of wireless standards that are commonly used for local wireless communications for PCs and laptops called 802.11. Currently, 802.11b and 802.11a are two basic standards that are accepted on a wider scale today. These standards are transmitted by using electromagnetic waves. Wireless signals as a whole can either be radio frequency (RF) or infrared frequency (IR), both being part of the electromagnetic spectrum (Boncella, 2002). Infrared (IR) broadcasting is used for close range communication and is specified in IEEE 802.11. The IR 802.11 implementation is based on diffuse IR which reflects signals off surfaces such as a ceiling and can only be used indoors. This type of transport is seldom used.

The most common physical transport is RF. The 802.11 standard uses this transport. Of the RF spectrum, the 802.11 standard uses the Industrial, Scientific, and Medical (ISM) RF band. The ISM band is designated through the following breakdown:

- The I-Band (from 902 MHz to 928MHz)
- The S-Band (from 2.4GHz to 2.48GHz)
- The M-Band (from 5.725GHz to 5.85GHz)

802.11b is the most accepted standard in wireless LANs (WLANs). This specification operates in the 2.4 gigahertz (GHz) S-band and is also known as wireless fidelity (WiFi). The speeds at which 802.11b can have data transfer rates is a maximum of 11 megabits per second (Boncella, 2002).

The 802.11a standard, commonly called WiFi5, is also used and operates with minor differences from the 802.11b standard. It operates in the M-band at 5.72GHz. The amount of data transfer has been greatly increased in this standard. The max link rate is 54 Mbps (Boncella, 2002).

There are other variations of 802.11 that may be used on a wider basis very soon. These are 802.11g and 802.11i. 802.11g operates in the same 2.4GHz S-band as 802.11b. Because they operate in the same band, 802.11g is compatible with 802.11b. The difference is that 802.11g is capable of a max link rate of 54 Mbps. The 802.11i standard is supposed to improve on the security of the Wired Equivalent Privacy (WEP) encryption protocol (Boncella, 2002).

The future of the 802.11 standard will bring other specifications—802.11c "helps improve interoperability between devices," 802.11d "improves roaming," 802.11e "is touted for its quality of service," 802.11f "regulates inter-access-point handoffs," and 802.11h "improves the 5GHz spectrum" (Worthen, 2003).

Another option for close range communication between devices is Bluetooth technology or through infrared port usage. Bluetooth is a short-range wireless standard that allows various devices to communicate with one another in close proximity, up to 10 meters (Tarasewich et al., 2002). The Infrared Data Association (IrDA) developed a personal area network standard based on infrared links, in 1994, which brought technology that is extremely useful in transferring applications and data from handheld devices such as PDAs (Agrawal et al., 2003) and between computers and other peripheral devices. It requires line of sight and covers a shorter distance than Bluetooth.

WLAN Architecture

A WLAN architecture is built from stations and an access point (AP). The basic structure of a WLAN is the Basic Service Set (BSS). A BSS may either be an independent BSS or an infrastructure BSS. (Boncella, 2002, p. 271)

An independent BSS does not use access points. Instead, the stations communicate with each other directly. They do have to be within range for this to occur. These types of networks are called ad hoc WLANs. They are generally created for short periods of time for such examples as meetings where a network needs to be established (Boncella, 2002). Another option for close range communication between devices wirelessly is Bluetooth technology or through infrared port usage.

An infrastructure BSS uses access points to establish a network. Each station must be associated with an AP because all the communications that transpire between stations run through the APs. Some restricted access is established because of the need to be associated with an AP (Boncella, 2002).

An Extended Service Set (ESS) can be created by these BSSs. A backbone network is needed to connect the BSSs. The purpose of creating an ESS is so that a user will have what is called "transition mobility." "If a station has transition mobility, a user is able to roam from BSS to BSS and continue to be associated with a BSS and also have access to the backbone network with no loss of connectivity" (Boncella, 2002, p. 272).

THE DECISION PROCESS

Usage of the Decision Model

After analyzing all of the different technologies in the wireless arena, the first decision that has the most impact on the wireless solution selected is the coverage needed by the wireless technology. There are three basic coverage areas that separate the wireless solution. The first is very short range coverage—30 feet or less. If the coverage needed is this small, the immediate solution is to use either an infrared port or use Bluetooth technology.

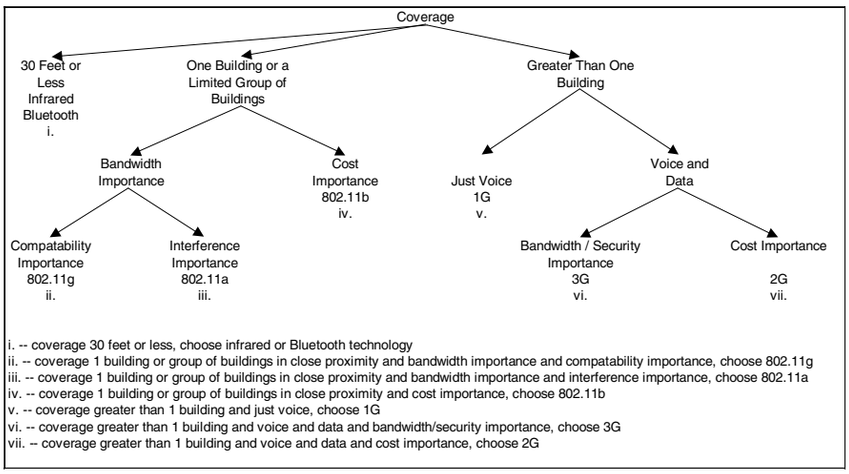
The second coverage area is larger than 30 feet, but is still somewhat concentrated. The solution for coverage that is needed just within one building or multiple buildings is a wireless LAN (WLAN). Because there are different solutions in the 802.11 standards, further analysis and breakdowns are needed. The second breakdown under this coverage area is a selection of what is more important between cost and amount of bandwidth. Because of the strong relationship between increased bandwidth and increased cost, these events are determined to be mutually exclusive. If keeping the cost down is more important then the solution is the 802.11b standard. If bandwidth is more important (due to a need for high data transfers), yet another breakdown occurs. The selection then depends on whether compatibility with other technologies is more important or if interference due to over saturation of the S-band is more important. These are deemed mutually exclusive because only two other 802.11 standards remain: 802.11g and 802.11a. The main difference is the band that is used. 802.11g uses the same S-band as 802.11b so there is compatibility for users of 802.11g with 802.11b APs, but at the same time, other devices such as cordless phones use the same band so interference can occur if the S-band is saturated or will become an issue in the future. If a more "clean" channel is desired with less interference the 802.11a standard is the appropriate solution. These two standards 802.11a and 802.11g both provide the same data rates.

The third and last coverage area is for distances that span farther than one building. If only voice is needed the solution is easy—a 1G technology would be the most cost efficient. Although this technology may become displaced by 2G or 3G technology, it is still an option in more remote areas which may have limited or no digital network service coverage. If voice and data services are needed, there are still two options, 2G and 3G. The main difference, again, is whether bandwidth or cost is more important. The difference in this breakdown, however, is that 3G provides an added level of security with device location. 3G also has higher bandwidth capabilities than 2G, so if bandwidth and an added level of security are more important than cost, a 3G technology should be chosen. If cost is more important, then 2G is the sufficient solution.

Since wireless networks lack the bandwidth of their wired counterparts, applications that run well on a wired network may encounter new problems when ported to a mobile environment (Tarasewich et al., 2002).

Although 3G has higher bandwidth capabilities and may provide an added level of security with device location, the cost of deploying the necessary technologies and security to 3G is greater than 2G, which may impact whether or not to implement new technology. Therefore, some companies are instead purchas-

Figure 1. Decision model


Coverage
30 Feet or Less Infrared Bluetooth i.
One Building or a Limited Group of Buildings
Greater Than One Building
Bandwidth Importance
Cost Importance 802.11b iv.
Just Voice 1G v.
Voice and Data
Compatability Importance 802.11g ii.
Interference Importance 802.11a iii.
Bandwidth / Security Importance 3G vi.
Cost Importance 2G vii.
i. -- coverage 30 feet or less, choose infrared or Bluetooth technology
ii. -- coverage 1 building or group of buildings in close proximity and bandwidth importance and compatability importance, choose 802.11g
iii. -- coverage 1 building or group of buildings in close proximity and bandwidth importance and interference importance, choose 802.11a
iv. -- coverage 1 building or group of buildings in close proximity and cost importance, choose 802.11b
v. -- coverage greater than 1 building and just voice, choose 1G
vi. -- coverage greater than 1 building and voice and data and bandwidth/security importance, choose 3G
vii. -- coverage greater than 1 building and voice and data and cost importance, choose 2G

ing data optimization software that can significantly increase data transmission speeds by using existing wireless connections (Tarasewich et al., 2002).

Limitations of Model

The limitation to using this decision model is that specific coverage areas for the different wireless generations' technologies were not taken into consideration. The reason that this coverage is a limitation is because of the many different carriers or providers of the technologies that exist and those providers having different coverage. Another limitation is the viewed context of using the model. The model focuses only on a "domestic" context as opposed to a global context. Demand for wireless applications differs around the world (Jarvenpaa et al., 2003; Tarasewich et al., 2002). In wireless technology, there are different standards used in the US as compared to other countries such as Europe and Asia.

TECHNOLOGY CONCERNS AND SECURITY ISSUES

Technology Concerns

There are several concerns for managers when investing in wireless technologies. One of the first concerns is that there is no single, universally accepted standard. This point raises questions or concerns over compatibility and interoperability between systems. Since standards may vary from country to country, it may become difficult for devices to interface with networks in different locations (Tarasewich et al., 2002). Thus, organizations may be hesitant to adopt a particular technology because they are afraid of high costs associated with potential obsolescence or incompatibility of technologies which they may decide to use.

Limitations of the technology are also an issue. Because many business applications take up considerable space and may access very large database files, the limitation of bandwidth could also be a concern. Even smaller files over a 2G device would take a long time to transfer. There are concerns regarding people and the change in culture that may need to take place. "Companies, employees, and consumers must be willing to change their behaviors to accommodate mobile capabilities." They may also have to "adapt their processes, policies, expectations and incentives accordingly" (Smith et al., 2002, p. 473).

Coverage area is also an issue. For example, a WLAN with numerous access points may need to be setup so that the mobile user can have access to the network regardless of user location and access point. Service providers of 1G, 2G, and / or 3G technologies may have areas that may not get service as well. The question whether seamless integration exists as far as working from a desktop or PC and then taking information to a mobile device such as a Personal Digital Assistant may also be a concern (Tarasewich et al., 2002).

Security is always an issue with wireless technology. Authentication is especially important in WLANs because of the wireless coverage boundary problems. No physical boundaries exist in a WLAN. Thus access to the systems from outside users is possible. Another concern is whether many devices are using the same frequency range. If this is the case, the devices may interfere with one another. Some of the interference is intentional because of "frequency hopping" which is done for security purposes (Tarasewich et al., 2002).

Because of capabilities to access wireless networks, data integrity is more of an issue than in wired networks. If data is seen at all and the information is confidential, there could be valuable information leaked that can be detrimental to a firm or organization (Smith et al., 2002). Viruses and physical hardware are sources of security issues as well. Mobile devices such as PDAs can be stolen from authorized users. Viruses can be sent wirelessly with the stolen device and then destroyed after the virus has been sent, thus making it difficult to identify the individual at hand (Tarasewich et al., 2002). With packet-switched services for mobile devices and with WLANs, the user of the devices has an "always-on" feature. The users are more susceptible to hacking when they are always on the wireless network (Smith et al., 2002).

WLAN Security Exploits

According to Robert Boncella, a number of security exploits exist related to wireless LANs. The first security exploit, an insertion attack, is when someone "inserts" themselves into a BSS that they are not suppose to have access to, usually to gain access to the Internet at no cost to them. A person can also "eavesdrop" by joining a BSS or setting up their own AP that may be able to establish itself as an AP on an infrastructure BSS. When the person has access, they can either run packet analysis tools or just analyze the traffic. Similarly, a person may try to clone an AP with the intent to take control of the BSS. If an AP is broadcasting because it is setup to act like a hub instead of a switch, monitoring can take place as well (Boncella, 2002).

A denial of service attack is one in which the usage of the wireless LAN is brought to a halt because of too much activity using the band frequencies. This can also happen by cloning a MAC or IP address. The effect is the same: access is brought to a halt. This is what is meant by a client-to-client attack. There are also programs that will attempt access to a device or program that requires a password and can be directed at an AP until access is granted—also known as a brute force attack against AP passwords. In WLANs, the protocol for encryption is Wired Equivalent Privacy (WEP), which is susceptible to compromise. The last exploit is misconfiguration. When a firm or organization gets an AP it usually ships with default settings (including default password, broadcasting, etc.) which if not changed can compromise the network since the knowledge of default settings is generally available to the public (Boncella, 2002).

Minimizing Security Issues

There are actions that can help reduce the security risks. Encryption technologies exist that can help ensure that data is not easily read. The problem with this is that developers of encryption protocols need to make them more efficient so bandwidth overhead is not a drain on the data rates that the individual will experience. Encryption is not always foolproof either.

Another method of reducing security issues uses information regarding device location to authenticate and allow access. Then, if the device is stolen, locating it might be possible, but also, if the device travels outside the accepted coverage area, access can be stopped. Usage of biometrics in devices and for authentication is another option. Biometrics that can be used would include thumbprint and/or retinal scanning ID devices (Tarasewich et al., 2002).

When firms or organizations decide or use WLAN technologies, there are three basic methods that can be used to secure the access to an AP: Service Set Identifier (SSID), Media Access Control (MAC) address filtering, and Wired Equivalent Privacy (WEP). One has to be careful with using SSID as a method of security, however, because it is minimal in nature and an AP can be setup to broadcast the SSID, which would then have no effect on enhancing security (Boncella, 2002).

The second method that can be used to help secure an AP is MAC address filtering. When used, only the stations with the appropriate MAC addresses can access the WLAN. The problem is that the MAC addresses have to be entered manually into the AP which can take significant time. Maintenance also can

be a hassle for larger firms because of the time it takes to keep the list up to date (Boncella, 2002).

The last method for WLAN security is usage of Wired Equivalent Privacy (WEP). The 802.11 specifications use WEP as the designated encryption method. It encrypts the data that is transferred to an AP and the information that is received from an AP (Boncella, 2002).

WEP is not totally secure, however. Programs exist that use scripts to attack a WLAN and WEP keys can be discovered. There may be a new solution which may replace WEP, called the Advance Encryption Standard (AES). Further development of 802.11 standards may also help alleviate some of the security vulnerabilities (Boncella, 2002).

Even though WEP is not completely secure and it does take up bandwidth resources, it is still recommended that it is used along with MAC address filtering and SSID segmentation. In WLANs, it is also recommended that clients password protect local drives, folders, and files. APs should be changed from their default settings and should not broadcast these SSIDs. If a firm or organization wants end-to-end security, the use of a Virtual Private Network (VPN) is possible. The technology has been established for quite some time and allows for users to use an "untrusted" network for secure communications. It is an increased cost and a VPN server and VPN client have to be used (Boncella, 2002).

CONCLUSION

While these different technology specifications are important in the decision making process because each of them are different and allow for different capabilities, it is also important to realize that decisions related to investing in technology such as modifications or restructuring to the business model can have an affect on investment. Also, investments in wireless technology can follow the investment options as well, thus potentially changing the path(s) using the decision model. Managers can use this decision model to plan their wireless technology implementation and applications.

The future of wireless technology may also bring more devices that can operate using the many different standards and it may be possible that a global standard is accepted such as the expected plans for the 3G technology UMTS.

Mobile and wireless technology has attracted significant attention among research and development communities. Many exciting research issues are being addressed and some are yet to be addressed and we hope that this article inspires others to do future research by expanding or enhancing this decision model. Researchers should need this decision model to categorize wireless technologies so that hypotheses and theories can be tested meaningfully.

Finally, this model should help information systems professionals to better identify meaningful wireless decision support systems (Power, 2004).

REFERENCES

Agrawal, M., Chari, K., & Sankar, R. (2003). Demystifying wireless technologies: Navigating through the wireless technology maze. *Communications of the Association for Information Systems, 12*(12), 166-182.

Anderson, J.E. & Schwager, P.H. (2004). SME adoption of wireless LAN technology: Applying the UTAUT model. *Proceedings of the 7th Annual Conference of the Southern Association for Information Systems, 1* (Vol. 1, pp. 39-43).

Boncella, R.J. (2002). Wireless security: An overview. *Communications of the Association for Information Systems, 9*(15), 269-282.

Byun, J. & Chatterjee, S. (2002). Network convergence: Where is the value? *Communications of the Association for Information Systems, 9*(27), 421-440.

Dunne, D. (2002). How to speak wireless. *CIO Magazine*. Retrieved April 12, 2003, from *http://www.cio.com/communications/edit/glossary.html*

Jarvenpaa, S.L., Lang, K., Reiner, T., Yoko, T., & Virpi, K. (2003). Mobile commerce at crossroads. *Communication of the ACM, 12*(46), 41-44.

Power, D.J. (2004). Specifying an expanded framework for classifying and describing decision support systems. *Communications of the Association for Information Systems, 13*(13), 158-166.

Smith, H., Kulatilaka, N., & Venkatramen, N. (2002). New developments in practice III: Riding the wave:

Extracting value from mobile technology. *Communications of the Association for Information Systems, 8*(32), 467-481.

Tarasewich, P. & Warkentin, M. (2002). Information everywhere. *Information Systems Management, 19*(1), 8-13.

Tarasewich, P., Nickerson, R.C., & Warkentin, Merrill. (2002). Issues in mobile e-commerce. *Communications of the Association for Information Systems, 8*(3), 41-64.

Varshney, U. & Jain, R. (2001). Issues in emerging 4G wireless networks. *Computer, 34*(6), 94-96.

Varshney, U. & Vetter, R. (2000). Emerging mobile and wireless networks. *Communication of the ACM, 43*(6), 73-81.

Varshney, U. & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks and Applications, 7*(3), 185-198.

Worthen, B. (2003). Easy as A,B,C,D,E,F,G,H and I. *CIO Magazine*. Retrieved April 12, 2003, from *http://www.cio.com/archive/010103/3.html*

Zhang, D. (2003). Delivery of personalized and adaptive content to mobile devices: A framework and enabling technology. *Communications of the Association for Information Systems, 12*(13), 183-202.

KEY TERMS

Authentication: Verification that one is who they say they are.

Bandwidth: Range of frequencies within a communication channel or capacity to carry data.

Bluetooth: Short-range wireless technology limited to less than 30 feet.

Encryption: Scrambling of data into an unreadable format as a security measure.

IP: Internet Protocol, which is network layer protocol of the TCP/IP protocol suite concerned with routing packets through a packet-switched network.

Packet: A package of data found at the network layer and contains source and destination address information as well as control information.

Protocol: Rules governing the communication and exchange of data across a network or internetworks.

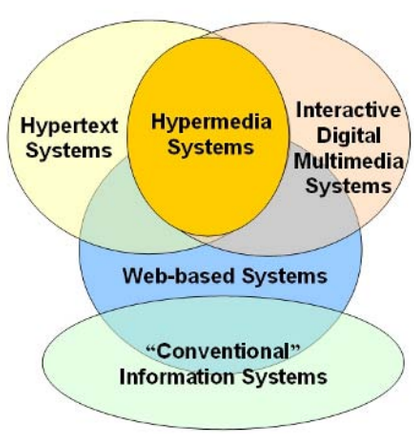
Designing Web-Based Hypermedia Systems

Michael Lang
National University of Ireland, Galway, Ireland

INTRODUCTION

Although its conceptual origins can be traced back a few decades (Bush, 1945), it is only recently that hypermedia has become popularized, principally through its ubiquitous incarnation as the World Wide Web (WWW). In its earlier forms, the Web could only properly be regarded a primitive, constrained hypermedia implementation (Bieber & Vitali, 1997). Through the emergence in recent years of standards such as eXtensible Markup Language (XML), XLink, Document Object Model (DOM), Synchronized Multimedia Integration Language (SMIL) and WebDAV, as well as additional functionality provided by the Common Gateway Interface (CGI), Java, plug-ins and middleware applications, the Web is now moving closer to an idealized hypermedia environment. Of course, not all hypermedia systems are Web based, nor can all Web-based systems be classified as hypermedia (see Figure 1). See the terms and definitions at the end of this article for clarification of intended meanings. The focus here shall be on hypermedia systems that are delivered and used via the platform of the WWW; that is, Web-based hypermedia systems.

Figure 1. Hypermedia systems and associated concepts



There has been much speculation that the design of Web-based hypermedia systems poses new or unique challenges not traditionally encountered within conventional information systems (IS) design. This article critically examines a number of issues frequently argued as being different—cognitive challenges of designing non-linear navigation mechanisms, complexity of technical architecture, pressures of accelerated development in "Web-time" environment, problems with requirements definition, the suitability of traditional design methods and techniques, and difficulties arising out of the multidisciplinary nature of hypermedia design teams. It is demonstrated that few of these issues are indeed new or unique, and clear analogies can be drawn with the traditions of conventional IS design and other related disciplines.

CRITICAL REVIEW OF PRINCIPAL DESIGN ISSUES AND CHALLENGES

Visualizing the Structure of Hypermedia Systems

Essentially, hypermedia attempts to emulate the intricate information access mechanisms of the human mind (Bush, 1945). Human memory operates by associating pieces of information with each other in complex, interwoven knowledge structures. Information is later recalled by traversing context-dependent associative trails. Hypermedia permits the partial mimicry of these processes by using hyperlinks to create non-linear structures whereby information can be associated and retrieved in different ways. Otherwise put, hypermedia facilitates multiple paths through a network of information where there may be many points of entry or exit. This especially is the case with Web-based hypermedia, where users can enter the system through a variety of side doors rather than through the front "home page". The undisciplined use of

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

hyperlinks can lead to chaotic "spaghetti code" structures (de Young, 1990). As systems scale up, this causes the substantial problem of "getting lost in cyberspace," whereby it becomes very difficult to locate information or navigate through the labyrinth of intertwined paths (Otter & Johnson, 2000; Thelwall, 2000).

Two principal reasons explain why difficulties in visualizing the structure of a hypermedia system may arise. First, non-linear navigation mechanisms lead to intricate multi-dimensional information architectures that are hard to conceptualize. Second, Web-based hypermedia systems are typically an amalgam of many different interconnected components, such as static Hypertext Markup Language (HTML) pages, client-side applets or scripts (e.g., Java, Javascript), dynamically generated pages (e.g., PHP, Perl, Active Server Pages, ColdFusion), media objects (e.g., JPEG, VRML, Flash, Quicktime) and back-end databases. Flows and dependencies are not as visible in Web-based hypermedia systems as they are for most conventional systems, and it can be quite difficult to form a clear integrated picture of the technical architecture (Carstensen & Vogelsang, 2001).

However, the phenomenon of systems being constructed using a multiplicity of components is not unique to Web-based hypermedia. In conventional systems design, tiered architectures that separate data, logic and interface layers are commonly used to assist seamless integration. One such approach is the Model-View-Controller (MVC) framework, which has also been found beneficial in Web-based hypermedia design (Izquierdo, Juan, López, Devis, Cueva & Acebal, 2003). Nor is the difficulty of designing non-linear navigation mechanisms unique to hypermedia. Within traditional printed media, certain types of material are intentionally designed to be used in a random-access non-linear manner, such as encyclopediae, thesauruses and reference works. According to Whitley (1998), hypermedia systems are different from other types of software applications because "the developers have to set up a number of alternatives for readers to explore rather than a single stream of text" (p. 70). This may be a new concept in software design, but elsewhere, technical writers have long experience of setting up multiple navigable paths in the design of electronic documentation, such as online help systems. It has been found that technical writing techniques can readily be adapted to navigation design for Web-based hypermedia systems (Eriksen, 2000).

Accelerated Development Environment

The capacity of organizations to respond and adapt quickly to rapidly changing environments is a well-recognised strategic issue. Accordingly, IS need to be flexible and able to adapt to changing business needs. Looking at trends in IS development over the past 20 years, project delivery times have dramatically shortened. In the early 1980s, Jenkins, Naumann and Wetherbe (1984) reported that the average project lasted 10.5 months. By the mid-1990s, the duration of typical projects had fallen to less than six months (Fitzgerald, 1997), and average delivery times for Web-based systems are now less than three months (Barry & Lang, 2003; Russo & Graham, 1999). These accelerated development cycles have given rise to the notion of "Web time" or "Internet speed" (Baskerville, Ramesh, Pries-Heje & Slaughter, 2003; O'Connell, 2001; Thomas, 1998), a development environment that is supposedly characterized by "headlong desperation and virtually impossible deadlines" (Constantine & Lockwood, 2002, p. 42).

Such compressed timeframes are made possible by the combined effect of two factors. First, modern-age, rapid-application development tools greatly speed up the development process, although it is sometimes argued that What-You-See-Is-What-You-Get (WYSIWYG) visual design tools invite a reckless "just-do-it" approach without much, if any, forethought. Second, the Web is an immediate delivery medium which, unlike traditional IS and off-the-shelf software applications, is not impeded by production, distribution and installation delays. Web-based systems can be easily and quickly launched by developing functional front-end interfaces, powered by crude but effective back-end software, which can later be modified and enhanced in such a manner that end users may be oblivious to the whole process.

Again, however, this phenomenon of reduced cycle times is not specific to Web-based hypermedia design, for it also affects the design of conventional systems (Kurata, 2001). Yourdon (1997) defined "death march" projects as those for which the normal parameters of time and resources are re-

duced by a factor of one-half or more. Such scenarios are now common across software development in general, not just Web-based systems. This is reflected by the growing interest amongst the general community of software developers in high-speed approaches such as agile methods, Rapid Application Development (RAD), timeboxing and commercial off-the-shelf (COTS) application packages. Indeed, one could say that this trend towards shorter cycles is reflective of a greater urgency in business today, brought about by dramatic advances in technology and exemplified by practices such as just-in-time (JIT) and business process re-engineering (BPR). Rapid flexible product development is a prerogative of the modern age (Iansiti & MacCormack, 1997). Considered thus, the phenomenon of "Web time" is not unique to Web-based hypermedia design, and it ought be regarded as an inevitable reality arising out of the age-old commercial imperative to devise faster, more efficient ways of working.

Requirements Elicitation and Verification

Traditionally, IS have served internal functions within organizations (Grudin, 1991). In contrast, the Web has an external focus—it was designed as a public information system to support collaborative work amongst distributed teams (Berners-Lee, 1996). As traditional IS are ported to the Web, they are turning inside-out and taking on a new focus, where brand consciousness and user experience design become important issues. In a sense, Web-based systems are shop windows to the world. Russo and Graham (1999) make the point that Web applications differ from traditional information systems because the users of Web applications are likely to be outside of the organization, and typically cannot be identified or included in the development process.

It is plainly true that for most Web-based systems, with the obvious exception of intranets, end users are external to the organization. Collecting requirements from a virtual population is difficult, and the same requirements elicitation and verification techniques that have traditionally been used in software systems design cannot be easily applied, if at all (Lazar, Hanst, Buchwalter & Preece, 2000). Although this is new territory for IS developers, the notion of a virtual population is quite typical for mass-market off-the-shelf software production and new product development (Grudin, 1991). In such situations, the marketing department fulfils a vital role as the voice of the customer. For example, Tognazzini (1995) describes how a team of designers, engineers and human factors specialists used scenarios to define requirements based on an understanding of the profiles of target users as communicated by marketing staff. Thus, marketing research techniques can be used in conjunction with user-centred requirements definition techniques to understand the requirements of a virtual population. To verify requirements, because end users can't readily be observed or interviewed, techniques such as Web log analysis and click tracking are useful (Lane & Koronois, 2001). The use of design patterns—tried and tested solutions to recurring design problems—is also advantageous (Lyardet, Rossi & Schwabe, 1999).

Applicability of Traditional Methods and Techniques

It is often argued that approaches and methods from traditional systems design are inappropriate for Web-based systems (Russo & Graham, 1999; Siau & Rossi, 2001). Murugesan, Deshpande, Hansen and Ginige (1999) speak of "a pressing need for disciplined approaches and new methods and tools," taking into account "the unique features of the new medium" (p. 2). It is arguable whether many of the features of Web-based hypermedia are indeed unique. Merely because an application is based on new technologies, its design should not necessarily require an altogether new or different approach. It may well be true that traditional methods and techniques are ill-suited to hypermedia design. However, for the same reasons, those methods can be argued to be inappropriate for conventional systems design in the modern age (Fitzgerald, 2000). Modern approaches, methods and techniques—such as rapid prototyping, incremental development, agile methods, use cases, class diagrams, graphic user interface (GUI) schematics and interaction diagrams—are arguably just as applicable to hypermedia design as to conventional systems design. Methods and techniques from other relevant disciplines such as graphic design and media production also bear examination, as evi-

denced by the findings of Barry and Lang (2003).

Diagrammatic models are often useful in systems design to help overcome the cognitive difficulties of understanding complex, abstract structures. It has been argued that diagramming techniques from traditional systems design are inappropriate for modelling hypermedia systems (Russo & Graham, 1999; Siau & Rossi, 2001). One could just as easily argue that the flow of control in modern visual event-driven and object-oriented programming languages (e.g., Microsoft Visual Basic, Borland Delphi, Macromedia Lingo) is such that traditional techniques such as structured flowcharts and Jackson Structured Programming (JSP) are of limited use. For these types of applications, modern techniques such as Unified Modelling Language (UML) are being used, as well as approaches inherited from traditional dynamic media (e.g., storyboarding). Both storyboarding and UML can likewise be applied to hypermedia design; indeed, a number of UML variants have been proposed specifically for modelling hypermedia systems (Baumeister, Koch & Mandel, 1999; Conallen, 2000).

Multidisciplinary Design Teams

Perhaps the only aspect of Web-based hypermedia systems design that is radically different from conventional systems design is the composition of design teams. In conventional systems development, designers tend to be primarily "computer professionals." This is not the case in hypermedia systems design, where team members come from a broad diversity of professional backgrounds, many of them non-technical. The challenge of managing communication and collaboration within multidisciplinary design teams is by no means trivial, and if mismanaged is potentially disastrous. Experiences reveal that discrepancies in the backgrounds of team members can give rise to significant communication and collaboration problems (Carstensen & Vogelsang, 2001). The multidisciplinary nature of design teams must be acknowledged in devising mechanisms to overcome the challenges of Web-based hypermedia design. Integrated working procedures, design approaches, diagramming techniques, toolset selection and mechanisms for specifying and managing requirements must all take this central aspect into consideration. The two foremost disciplines of Web-based hypermedia design are software engineering and graphic design (Lang, 2003), but alarmingly, it has been observed that these two factions have quite different value systems (Gallagher & Webb, 1997) and "appear to operate in distinctly different worlds" (Vertelney, Arent & Lieberman, 1990, p. 45). This is a considerable challenge which, if not addressed, could foil a project. Lessons can be learned from other disciplines that have successfully balanced the relationship between critical functionality and aesthetic attractiveness, such as architecture/civil engineering, automobile design and computer game development.

CONCLUSION

Throughout the history of computer systems design, it has been common amongst both researchers and practitioners to greet the arrival of much-hyped next-generation technologies by hailing them as profound advances that warrant entirely new approaches. Web/hypermedia design is another such example. However, Nielsen (1997) has commented that "software design is a complex craft and we sometimes arrogantly think that all its problems are new and unique" (p. 98). As this article reveals, few of the challenges of Web-based hypermedia design are indeed new or unique. Parallels can be drawn with lessons and experiences across a variety of disciplines, yet much of the literature on hypermedia design fails to appreciate the wealth of this legacy. Design methods, approaches and techniques can be inherited from many root disciplines, including traditional IS development, software engineering, human-computer interaction (HCI), graphic design, visual communications, marketing, technical writing, library information science, media production, architecture and industrial design. To paraphrase a well-known saying, those who choose not to draw from the well of cumulative knowledge are bound to foolishly repeat mistakes and to wastefully spend time seeking solutions where they might already exist. This article, therefore, concludes with a petition to hypermedia design researchers that they resist the temptation to dub themselves a "new" discipline (Murugesan et al., 1999), and instead reach out to explore the past and present experiences of related traditions.

REFERENCES

Barry, C., & Lang, M. (2003). A comparison of "traditional" and multimedia information systems development practices. *Information and Software Technology, 45*(4), 217-227.

Baskerville, R., Ramesh, B., Pries-Heje, J., & Slaughter, S. (2003). Is Internet-speed software development different? *IEEE Software, 20*(6), 70-77.

Baumeister, H., Koch, N., & Mandel, L. (1999, October 28-30). Towards a UML extension for hypermedia design. In R.B. France & B. Rumpe (Eds.), *UML'99: The Unified Modeling Language - Beyond the Standard, Second International Conference, Fort Collins, CO Proceedings. Lecture Notes in Computer Science 1723* (pp. 614-629).

Berners-Lee, T. (1996). WWW: Past, present, and future. *IEEE Computer, 29*(10), 69-77.

Bieber, M., & Vitali, F. (1997). Toward support for hypermedia on the World Wide Web. *IEEE Computer, 30*(1), 62-70.

Bush, V. (1945). As we may think. *The Atlantic Monthly, 176*(1), 101-108.

Carstensen, P.H., & Vogelsang, L. (2001, June 27-29). Design of Web-based Information Systems – New Challenges for Systems Development? Paper presented at the *Proceedings of 9th European Conference on Information Systems (ECIS),* Bled, Slovenia.

Conallen, J. (2000). *Building Web applications with UML*. Reading, MA: Addison Wesley.

Constantine, L.L., & Lockwood, L.A.D. (2002). Usage-centered engineering for Web applications. *IEEE Software, 19*(2), 42-50.

de Young, L. (1990). Linking considered harmful. *Hypertext: Concepts, systems and applications* (pp. 238-249). Cambridge: Cambridge University Press.

Eriksen, L.B. (2000, June 9-11). Limitations and opportunities for system development methods in Web information system design. In R. Baskerville, J. Stage & J.I. DeGross (Eds.), *Organizational and Social Perspectives on Information Technology, IFIP TC8 WG8.2 International Working Conference on the Social and Organizational Perspective on Research and Practice in Information Technology,* Aalborg, Denmark (pp. 473-486). Boston: Kluwer.

Fitzgerald, B. (1997). The use of systems development methodologies in practice: A field study. *Information Systems Journal, 7*(3), 201-212.

Fitzgerald, B. (2000). Systems development methodologies: The problem of tenses. *Information Technology & People, 13*(3), 174-185.

Gallagher, S., & Webb, B. (1997, June 19-21). Competing paradigms in multimedia systems development: Who shall be the aristocracy? Paper presented at the *Proceedings of 5th European Conference on Information Systems (ECIS),* Cork, Ireland.

Grudin, J. (1991). Interactive systems: Bridging the gaps between developers and users. *IEEE Computer, 24*(4), 59-69.

Iansiti, M., & MacCormack, A. (1997). Developing products on Internet time. *Harvard Business Review, 75*(5), 108-117.

Izquierdo, R., Juan, A., López, B., Devis, R., Cueva, J.M., & Acebal, C.F. (2003, July 14-18). Experiences in Web site development with multidisciplinary teams. From XML to JST. In J.M.C. Lovelle, B.M.G. Rodríguez & M.D.P.P. Ruiz (Eds.), *Web engineering: International Conference, ICWE2003, Oviedo, Spain* (pp. 459-462). Berlin: Springer.

Jenkins, M.A., Naumann, J.D., & Wetherbe, J.C. (1984). Empirical investigation of systems development practices and results. *Information & Management, 7*(2), 73-82.

Kurata, D. (2001). Do OO in "Web time." *Visual Basic Programmer's Journal, 11*(1), 70.

Lane, M.S., & Koronois, A. (2001). A balanced approach to capturing user requirements in business-to-consumer Web information systems. *Australian Journal of Information Systems, 9*(1), 61-69.

Lang, M. (2003). Hypermedia systems development: A comparative study of software engineers

and graphic designers. *Communications of the AIS, 12*(16), 242-257.

Lazar, J., Hanst, E., Buchwalter, J., & Preece, J. (2000). Collecting user requirements in a virtual population: A case study. *WebNet Journal, 2*(4), 20-27.

Lyardet, F., Rossi, G., & Schwabe, D. (1999). Discovering and using design patterns in the WWW. *Multimedia Tools and Applications, 8*(3), 293-308.

Murugesan, S., Deshpande, Y., Hansen, S., & Ginige, A. (1999, May 16-17). *Web engineering: A new discipline for development of Web-based systems.* Paper presented at the proceedings of 1st ICSE Workshop on Web Engineering, Los Angeles, CA.

Nielsen, J. (1997). Learning from the real world. *IEEE Software, 14*(4), 98-99.

O'Connell, F. (2001). *How to run successful projects in Web time*. London: Artech House.

Otter, M., & Johnson, H. (2000). Lost in hyperspace: metrics and mental models. *Interacting with Computers, 13*(1), 1-40.

Russo, N.L., & Graham, B.R. (1999). A first step in developing a Web application design Methodology: Understanding the environment. In A.T. Wood-Harper, N. Jayaratna & J.R.G. Wood (Eds.), *Methodologies for developing and managing emerging technology based information systems: Proceedings of 6th International BCS Information Systems Methodologies Conference* (pp. 24-33). London: Springer.

Siau, K., & Rossi, M. (2001). Information modeling in the Internet age - Challenges, issues and research directions. In M. Rossi & K. Siau (Eds.), *Information modeling in the new millennium* (pp. 1-8). Hershey: Idea Group Publishing.

Thelwall, M. (2000). Commercial Web sites: lost in cyberspace? *Internet Research, 10*(2), 150-159.

Thomas, D. (1998, October). Web time software development. *Software Development Magazine,* 78-80.

Tognazzini, B. (1995). *Tog on software design*. Reading, MA: Addison Wesley.

Vertelney, L., Arent, M., & Lieberman, H. (1990). Two disciplines in search of an interface: Reflections on a design problem. In B. Laurel (Ed.), *The art of human-computer interface design* (pp. 45-55). Reading, MA: Addison Wesley.

Whitley, E.A. (1998, December 13-16). *Methodism in practice: Investigating the relationship between method and understanding in Web page design.* Paper presented at the proceedings of 19th International Conference on Information Systems (ICIS), Helsinki, Finland.

Yourdon, E. (1997). *Death march: The complete software developer's guide to surviving "Mission Impossible" projects*. Upper Saddle River, NJ: Prentice Hall.

KEY TERMS

Commercial Off-the-Shelf (COTS) applications: An approach to software development where, instead of attempting to build an application from scratch, a generic standardized package is purchased that contains all the main functionality. This package is then configured and customized so as to meet the additional specific requirements.

Hypermedia: "Hypermedia" is often taken as synonomous with "hypertext," though some authors use "hypermedia" to refer to hypertext systems that contain not just text data, but also graphics, animation, video, audio and other media. Principal defining features of a hypermedia system are a highly interactive, visual, media-rich user interface and flexible navigation mechanisms. Hypermedia is a specialized type of interactive digital multimedia.

Hypertext: An approach to information management in which data is stored as a network of inter-related nodes (also commonly known as "documents" or "pages") that may be purposefully navigated or casually browsed in a non-linear sequence by means of various user-selected paths, following hyperlinks. These hyperlinks may be hard-coded into the system or dynamically generated at runtime.

Incremental Development: An approach to software development in which fully working versions of a system are successively delivered over time, each new increment (version) adding to and upgrading the functionality of the previous version. May be used in conjunction with "timeboxing," whereby a "wish list" of requirements is prioritized and ordered into a staged plan of increments to be rolled out over time.

Interactive Digital Multimedia: *Interactive* digital multimedia systems enable end users to customize and select the information they see and receive by actively engaging with the system (e.g., tourism kiosk, interactive television), as opposed to *passive* multimedia where the end user has no control over the timing, sequence or content (e.g., videotape, linear presentation) (see also multimedia).

JPEG: A standard file type for computerized images, determined by the Joint Photographic Experts Group.

Multimedia: Broadly defined, multimedia is the blending of sound, music, images and other media into a synchronized whole. Such a definition is perhaps too wide, for it may be taken to include artistic works, audiovisual presentations, cinema, theatre, analogue television and other such media forms. A more precise term is "digital multimedia," meaning the computer-controlled integration of text, graphics, still and moving images, animation, sounds and any other medium where every type of information can be represented, stored, transmitted and processed digitally. (See also interactive digital multimedia).

Rapid Application Development (RAD): RAD is a software development approach that aims to enable speedier development, improve the quality of software and decrease the cost of development. It emphasizes the use of computer-aided software engineering (CASE) tools and fourth-generation programming languages (4GLs) by highly-trained developers, and uses intensive workshops to assist requirements definition and systems design.

VRML: Virtual Reality Markup Language, used to model three-dimensional worlds and data sets on the Internet.

Web-Based Systems: A loose term that in its broadest sense embraces all software systems that somehow rely upon the WWW as a platform for execution, including not just interactive Web sites but also applications such as Web crawlers and middleware. In a narrow sense, it is generally taken to mean systems for which human-computer interaction is mediated through a Web browser interface.

WYSIWYG Visual Design Tools: A category of application development tools that emphasizes the visual design of the front-end graphical user interface (GUI); that is, What You See Is What You Get (WYSIWYG). These tools often have prototyping features, such as automatic code generation and customizable in-built application templates. Examples include Microsoft Frontpage and Macromedia Dreamweaver.

Digital Filters

Gordana Jovanovic-Dolecek
INAOE, Mexico

INTRODUCTION

A signal is defined as any physical quantity that varies with changes of one or more independent variables, and each can be any physical value, such as time, distance, position, temperature, or pressure (Oppenheim & Schafer, 1999; Elali, 2003; Smith, 2002). The independent variable is usually referred to as "time". Examples of signals that we frequently encounter are speech, music, picture, and video signals. If the independent variable is continuous, the signal is called continuous-time signal or analog signal, and is mathematically denoted as $x(t)$. For discrete-time signals the independent variable is a discrete variable and therefore a discrete-time signal is defined as a function of an independent variable n, where n is an integer. Consequently, $x(n)$ represents a sequence of values, some of which can be zeros, for each value of integer n. The discrete–time signal is not defined at instants between integers and is incorrect to say that $x(n)$ is zero at times between integers. The amplitude of both the continuous and discrete-time signals may be continuous or discrete. Digital signals are discrete-time signals for which the amplitude is discrete. Figure 1 illustrates the analog and the discrete-time signals.

Most signals we encounter are generated by natural means. However, a signal can also be generated synthetically or by computer simulation (Mitra, 2001).

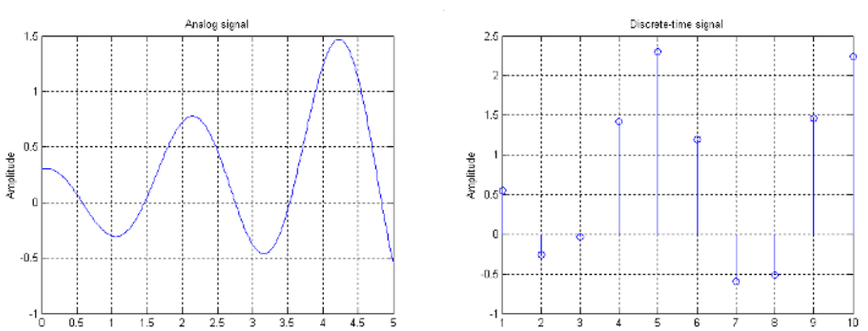
A signal carries information, and objective of signal processing is to extract useful information carried by the signal. The method of information extraction depends on the type of signal and the nature of the information being carried by the signal. "Thus, roughly speaking, signal processing is concerned with the mathematical representation of the signal and algorithmic operation carried out on it to extract the information present" (Mitra, 2001, p. 1).

Analog signal processing (ASP) works with the analog signals, while digital signal processing (DSP) works with digital signals. Since most of the signals we encounter in nature are analog, DSP consists of these three steps:

- A/D conversion (transformation of the analog signal into the digital form)
- Processing of the digital version
- Conversion of the processed digital signal back into an analog form (D/A)

We now mention some of the advantages of DSP over ASP (Diniz, Silva, & Netto, 2002; Grover & Deller, 1999; Ifeachor & Jervis, 2001; Mitra, 2001; Stein, 2000):

Figure 1. Examples of analog and discrete-time signals



Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

- Less sensitivity to tolerances of component values and independence of temperature, aging and many other parameters.
- Programmability, that is, the possibility to design one hardware configuration that can be programmed to perform a very wide variety of signal processing tasks simply by loading in different software.
- Several valuable signal processing techniques that cannot be performed by analog systems, such as for example linear phase filters.
- More efficient data compression (maximum of information transferred in the minimum of time).
- Any desirable accuracy can be achieved by simply increasing the word length.
- Applicability of digital processing to very low frequency signals, such as those occurring in seismic applications. (Analog processor would be physically very large in size.)
- Recent advances in very large scale integrated (VLSI) circuits, make possible to integrate highly sophisticated and complex digital signal processing systems on a single chip.

Nonetheless, DSP has some disadvantages (Diniz et al., 2002; Grover & Deller, 1999; Ifeachor & Jervis, 2001; Mitra, 2001; Stein, 2000):

- Increased complexity: The need for additional pre-and post-processing devices such as A/D and D/A converters and their associated filters and complex digital circuitry.
- The limited range of frequencies available for processing.
- Consummation of power: Digital systems are constructed using active devices that consume electrical power whereas a variety of analog processing algorithms can be implemented using passive circuits employing inductors, capacitors, and resistors that do not need power.

In various applications, the aforementioned advantages by far outweigh the disadvantages and with the continuing decrease in the cost of digital processor hardware, the field of digital signal processing is developing fast. "Digital signal processing is extremely useful in many areas, like image processing, multimedia systems, communication systems, audio signal processing" (Diniz et al., 2002, pp. 2-3).

Figure 2. Digital filter

The system which performs digital signal processing i.e., transforms an input sequence $x(n)$ into a desired output sequence $y(n)$, is called a digital filter (see Figure 2).

We consider a filter is linear-time invariant system (LTI). The linearity means that the output of a scaled sum of the inputs is the scaled sum of the corresponding outputs, known as the principle of superposition. The time invariance says that a delay of the input signal results in the same delay of the output signal.

TIME-DOMAIN DESCRIPTION

If the input sequence $x(n)$ is a unit impulse sequence $\delta(n)$ (Figure 3),

$$\delta(n) = \begin{cases} 1 & for \quad n = 0 \\ 0 & otherwise \end{cases}, \tag{1}$$

then the output signal represents the characteristics of the filter called the impulse response, and denoted by $h(n)$. We can therefore describe any digital filter by its impulse response $h(n)$.

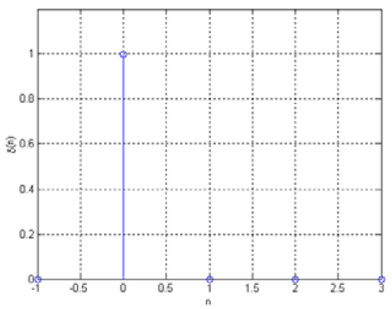
Depending on the length of the impulse response $h(n)$, digital filters are divided into filters with the *finite impulse response* (FIR) and *infinite impulse response* (IIR).

For example, let us consider an FIR filter of length $N = 8$ and impulse response as shown in Figure 4a.

$$h(n) = \begin{cases} 1/8 & for \quad 0 \le n \le 7 \\ 0 & otherwise \end{cases}, \tag{2}$$

In Figure 4b, the initial 20 samples of the impulse response of the IIR filter

Figure 3. Unit impulse sequence



$$h(n) = \begin{cases} 0.8^n & for \quad 0 \le n \\ 0 & for \quad n < 0 \end{cases}. \tag{3}$$

are plotted.

In practical applications, one is only interested in designing stable digital filters, that is, whose outputs do not become infinite. The stability of a digital filter can be expressed in terms of the absolute values of its unit sample responses (Kuc, 1988; Mitra, 2001; Proakis & Manolakis, 1996; Smith, 2002).

$$\sum_{n=-\infty}^{\infty} |h(n)| < \infty . \tag{4}$$

Because the summation (4) for an FIR filter is always finite, FIR filters are always stable. Therefore, the stability problem is relevant in designing IIR filters.

The operation in time domain which relates the input signal $x(n)$, impulse response $h(n)$ and the output signal $y(n)$, is called the *convolution*, and is defined as

$$y(n) = x(n) * h(n) = h(n) * x(n) =$$

$$\sum_k h(k)x(n-k) = \sum_k x(k)h(n-k) \tag{5}$$

where * is the standard sign for convolution. Figure 5 illustrates the convolution operation.

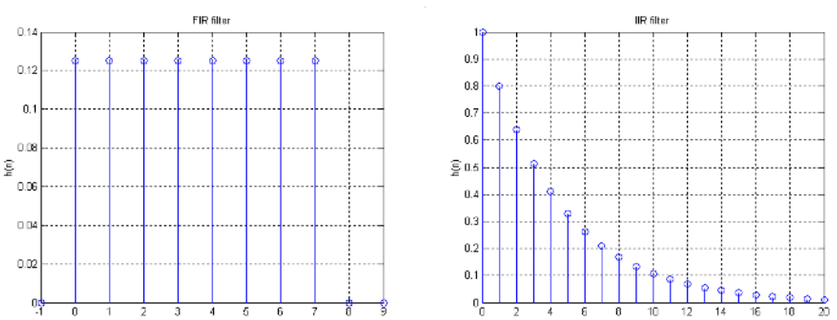
The output $y(n)$ can also be computed recursively using the following difference equation (Kuc, 1988; Mitra, 2001; Proakis & Manolakis, 1996; Silva & Jovanovic-Dolecek, 1999)

$$y(n) = \sum_{k=0}^{M} b_k x(n-k) + \sum_{k=1}^{N} a_k y(n-k), \tag{6}$$

where $x(n\text{-}k)$ and $y(n\text{-}k)$ are input and output sequences $x(n)$ and $y(n)$ delayed by k samples, and b_k and a_k, are constants. The order of the filter is given by the maximum value of N and M. The first sum is a *non-recursive*, while the second sum is a *recursive* part. Typically, FIR filters have only non-recursive part, while IIR filters always have the recursive part. As a consequence, FIR and IIR filters are also known as non-recursive and recursive filters, respectively.

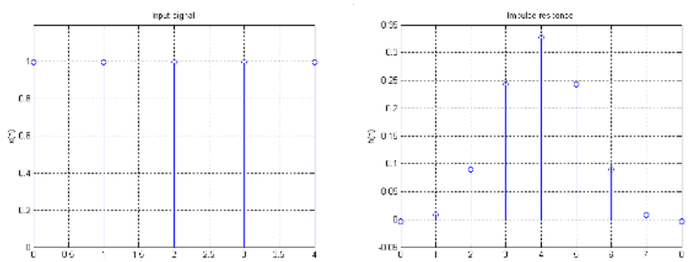
From (6) we see that the principal operations in a digital filter are multiplications, delays and additions. From the equation (6) we can draw the structure of the digital filter which is also known as a direct form and is shown in Figure 6. More details about filter structures can be found for example in Mitra (2001), Kuc (1988), and Proakis and Manolakis (1996).

Figure 4. Impulse responses of an FIR and IIR filter

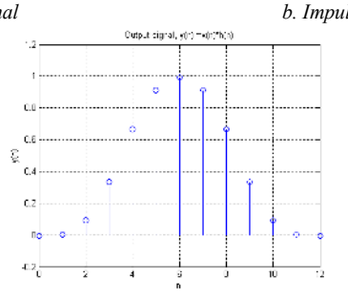


a. FIR filter *b. IIR filter*

Figure 5. Convolution



a. Input signal *b. Impulse response*



c. Output signal

DIGITAL FILTERS IN THE TRANSFORM DOMAIN

The popularity of the transform domain in DSP is due to the fact that more complicated time domain operations are converted to much simpler operations in the transform domain. Moreover, different characteristics of signals and systems can be better observed in the transform domain (Mitra 2001; Smith, 2002). The representation of digital filters in the transform domain is obtained using the Fourier transform and z- transform.

The Fourier transform of the signal $x(n)$ is defined as

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{j\omega n}, \quad (7)$$

where ω is digital frequency in radians and e^{jwn} is an exponential sequence. In general case, the Fourier transform is a complex quantity.

The convolution operation becomes multiplication in the frequency domain,

$$Y(e^{j\omega}) = X(e^{j\omega})H(e^{j\omega}), \quad (8)$$

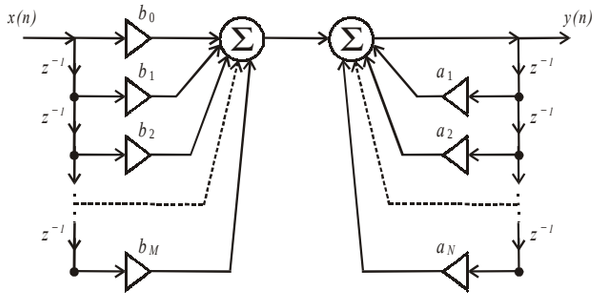
where $Y(e^{j\omega})$, $X(e^{j\omega})$, and $H(e^{j\omega})$, are Fourier transforms of $y(n)$, $x(n)$ and $h(n)$, respectively. The quantity $H(e^{j\omega})$ is called the *frequency response* of the digital filter, and it is a complex function of the frequency ω with a period 2π. It can be expressed in terms of its real and imaginary parts, $H_R(e^{j\omega})$ and $H_I(e^{j\omega})$, or in terms of its magnitude $\left|H(e^{j\omega})\right|$ and phase $\phi(\omega)$

$$H(e^{j\omega}) = H_R(e^{j\omega}) + jH_I(e^{j\omega}) = \left|H(e^{j\omega})\right|e^{j\phi(\omega)} \quad (9)$$

The amplitude $\left|H(e^{j\omega})\right|$ is called the magnitude response and the phase $\phi(\omega)$ is called the phase response of the digital filter. For a real impulse response digital filter, the magnitude response is a real even function of ω, while the phase response is a real odd function of ω. Figure 7 illustrates the magnitude responses of $x(n)$, $y(n)$, and $h(n)$, (previously shown in Figure 4).

In some applications, the magnitude response is expressed in the logarithmic form in decibels as

Figure 6. Direct form structure



$$G(\omega) = 20\log_{10}\left|H(e^{j\omega})\right| \quad \text{dB}\,, \tag{10}$$

where $G(\omega)$ is called the Gain function. Figure 8 illustrates Gain function of the filter $h(n)$ from Figure 5b.

z-transform is a generalization of the Fourier transform that allows us to use transform techniques for signals not having Fourier transform (Kuc, 1988; Proakis & Manolakis, 1999).

For the sequence $x(n)$, z-transform is defined as

$$X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n}\,, \tag{11}$$

where z is a complex variable. All values of z for which (11) converges are called the region of convergence (ROC).

z-transform of the unit sample response $h(n)$, denoted as $H(z)$, is called *system function*. Using z-transform of the Equation (6) we arrive at

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \quad . \tag{12}$$

For the FIR filter, all coefficients a_k, are zero, and consequently the denominator of the system function is simply 1. However, IIR filters always have the denominator different from 1.

Using the definition (11) we compute the z-transform of the FIR filter given in equation (2) as

$$H(z) = \frac{1}{8}\sum_{n=0}^{7} z^{-n} = \frac{1}{8}(1 + z^{-1} + \ldots + z^{-7}) = \frac{1}{8}\frac{1 - z^{-8}}{1 - z^{-1}}\,. \tag{13}$$

Similarly, the z-transform of the IIR filter given in (3) is

$$H(z) = \sum_{n=0}^{\infty} 0.8^n z^{-n} = \frac{1}{1 - 0.8z^{-1}}\,. \tag{14}$$

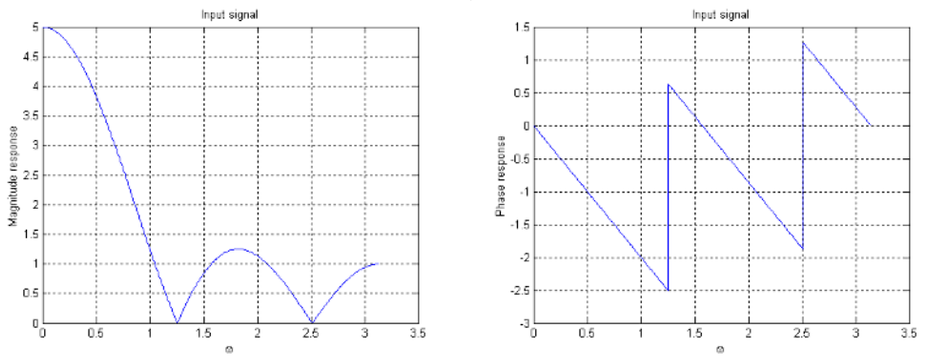
The equation (13) demonstrates that the FIR filter can also be expressed in a recursive form. Consequently, this filter is called recursive running–sum filter (RRS).

The roots of the numerator, or the values of z for which $H(z)$=0, define the locations of the zeros in the complex z plane. Similarly, the roots of the denominator, or the values of z for which $H(z)$ become infinite, define the locations of the poles. Both poles and zeros are called singularities. The plot of the singularities in z-plane is called the pole-zero pattern. A zero is usually denoted by a circle o and the pole by a cross x. An FIR filter has only zeros (poles are in the origin), whereas an IIR filter can have either both zeros and poles, or only poles, (zeros are in the origin). The filter is stable if all poles are inside the unit circle in z-plane. Figure 9 illustrates singularities in z-plane for filters (2) and (3). We can notice that both filters are stable. (FIR filter and IIR filter with the pole inside the unit circle, respectively.)

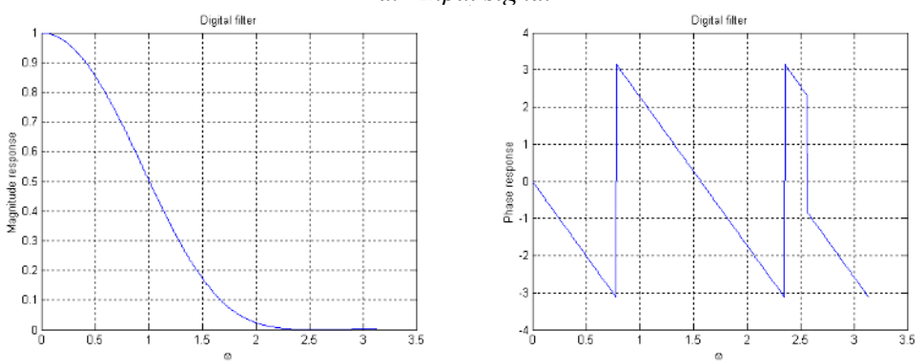
DESIGN OF DIGITAL FILTERS

The design of digital filter is the determination of a realizable system function $H(z)$ approximating the given frequency response specification (Mitra, 2001; Smith, 2002; Stearns, 2002; White, 2000). There are two major issues that need to be answered before one can develop $H(z)$. The first issue is the development of a reasonable magnitude specification from the requirements of the filter application. The second issue is the choice on whether an FIR or an IIR digital filter is to be designed (Mitra, 2001; White, 2000).

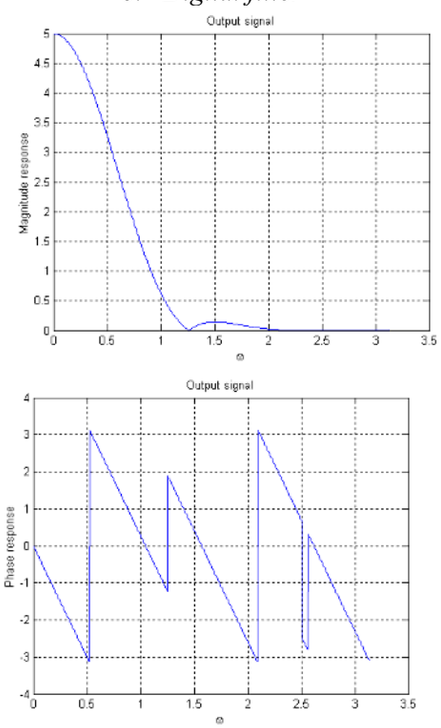
Figure 7. Magnitude and phase responses



a. Input signal

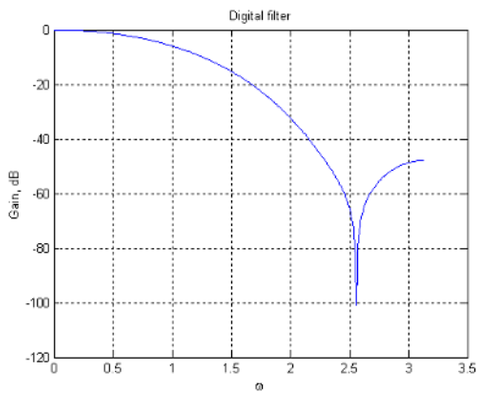


b. Digital filter



c. Output signal

Figure 8. Gain function of the digital filter



In most practical applications the problem of interest is the digital filter design for a given magnitude response specification. If necessary, the phase response of the designed filter can be corrected by equalizers filters (Mitra, 2001).

The filter that only passes low frequencies and rejects high frequencies is called a lowpass filter. The ideal lowpass filter has the magnitude specification givenby by

$$\left|H(e^{j\omega})\right| = \begin{cases} 1 & for \quad |\omega| \le \omega_c \\ 0 & for \quad \omega_c < |\omega| \le \pi \end{cases}, \tag{15}$$

where ω_c is called cutoff frequency. This filter can not be realized so the realizable specification is shown in Figure 10. The cutoff frequency ω_c is replaced by the transition band in which the magnitude specification is not given. The magnitude responses in the passband and the stopband are given with some acceptable tolerances, as shown in Figure 10.

The passband is defined for frequencies

$$0 \le \omega \le \omega_p, \tag{16}$$

where ω_p is called the passband edge frequency. The characteristic in the passband is defined as

$$1-\delta_1 \le \left|H(e^{j\omega})\right| \le 1+\delta_1 \qquad for |\omega| \le \omega_p \ , \tag{17}$$

where δ_1 is called the passband ripple.

In the stopband, defined by the stopband edge frequency ω_s

$$\omega_s \le \omega \le \pi \ , \tag{18}$$

the magnitude response approximates zero with an error of δ_2, called the stopband ripple

$$\left|H(e^{j\omega})\right| \le \delta_2 \qquad for \omega_s \le |\omega| \le \pi \ . \tag{19}$$

If the filter specification is given in terms of the Gain function in dB, the passband ripple *Rp* is then

$$R_p = 20\log_{10}(1-\delta_1) \ \text{dB}, \tag{20}$$

and the stopband attenuation A_s is

$$A_s = 20\log_{10}(\delta_2)\ \text{dB}. \tag{21}$$

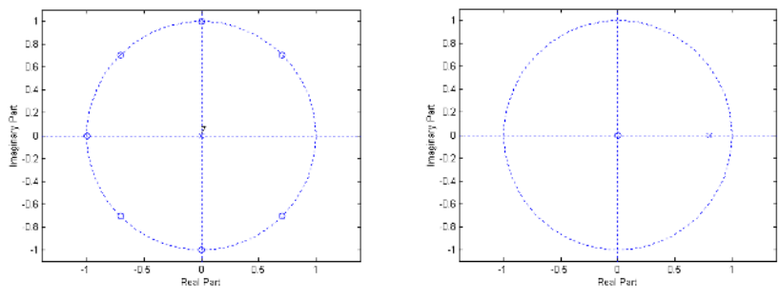
In a similar way, we can define the specifications for the highpass, bandpass, and stopband filters. For more details see White (2000) and Mitra (2001).

The principal methods for the design of FIR filters are: frequency sampling, window methods, weighted-least-squares (WLS), Remez method, and so on. For more details see White (2000) and Diniz et al. (2002).

As an example an FIR filter with the passband edge ω_p =0.2π, the stopband edge ω_p =0.3π, passband ripple R_p=0.1 dB and the stopband attenuation A_s=60 dB is plotted in Figure 11a and 11b. The designed filter has order of N = 57.

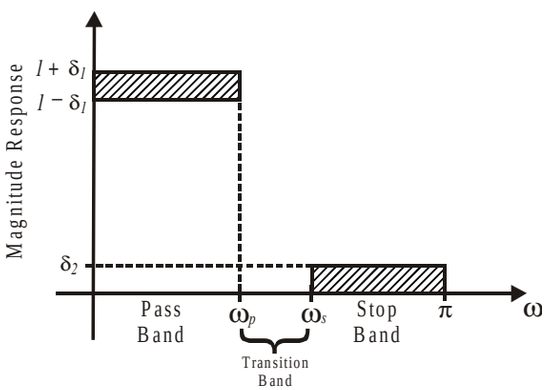
The most widely used methods for IIR filter design are extensions of the methods for the analog filter design (Mitra, 2001; Silva & Jovanovic-Dolecek, 1999; White, 2002). The reason is two-fold. Like IIR filters, analog filters have an infinite impulse response, and the methods for the design of analog filters are highly advanced. As a first step, the digital filter specification is converted into an analog lowpass filter specification, and an analog filter meeting this specification is designed. Next, the designed analog filter is transformed into a desired digital filter. Commonly used transformation methods are bilinear and impulse invariance method (Mitra, 2001, Silva & Jovanovic-Dolecek, 1999). If a filter other than a

Figure 9. Pole-zero pattern



a. FIR filter *b. IIR filter*

Figure 10. Lowpass filter specification



lowpass filter needs to be designed, the method also includes the frequency transformation in which the designed lowpass filter is transformed into the appropriate type (highpass, bandpass, or band-rejecting filter).

As an example, an elliptic filter is designed with the same specification as the filter plotted in Figure 11a and 11b. Its impulse response and the gain are plotted in Figure 11c and 11d. The impulse response, being infinite, is shown only for the initial 20 samples.

COMPARISON OF DIGITAL FILTERS

FIR filters are often preferred over IIR filters because they have many very desirable properties (Mitra, 2001; Proakis & Manolakis, 1996), such as linear phase, stability, absence of limit cycle, and good quantization properties. Arbitrary frequency responses can be designed, and excellent design techniques are available for a wide class of filters.

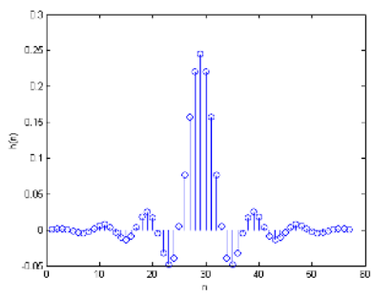
The main disadvantage of FIR filters is that they involve a higher degree of computational complexity compared to IIR filters with equivalent magnitude response. It has been shown that for most practical filter specifications the ratio of the FIR filter order and IIR filter order is typically of tens or more (Mitra, 2001), and as a result the IIR filter is computationally more efficient. However, if the linearity of the phase is required, the IIR filter must be equalized and in this case the savings in computation may no longer be that significant (Mitra, 2001).

In many applications where the linearity of the phase is not required, the IIR filters are preferable because of the lower computational requirements.

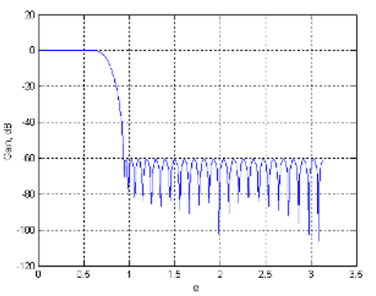
FIR filters of length N require $(N+1)/2$ multipliers if N is odd and $N/2$ multipliers if N is even, N-1 adders and N-1 delays. The complexity of the implementation increases with the increase in the number of multipliers.

Over the past few years, there have been a number of attempts to reduce the number of multipliers, like Adams and Willson (1983); Adams and Willson (1984); Ramakrishnan (1989); Bartolo, Clymer, Burgess, and Turnbull (1998) and so on. Another approach is a true multiplier-less design where the coefficients are reduced to simple integers or to simple combinations of powers of two, for example, Tai and Lin (1992), Yli-Kaakinen and Saramaki (2001), Liu, Chen, Shin, Lin, and Jou (2001), Coleman (2002), Jovanovic-Dolecek and Mitra (2002), and so on.

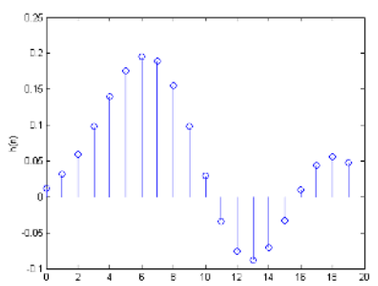
Figure 11. Design of the FIR and the IIR filters



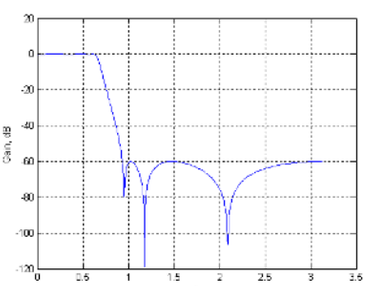
a. Impulse response of FIR filter



b. Gain in dB of designed FIR filter

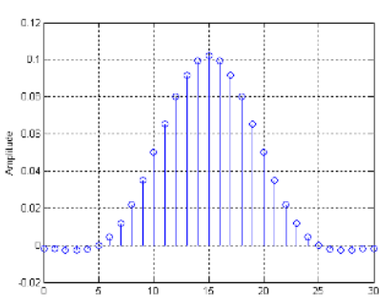


c. Impulse response of IIR filter

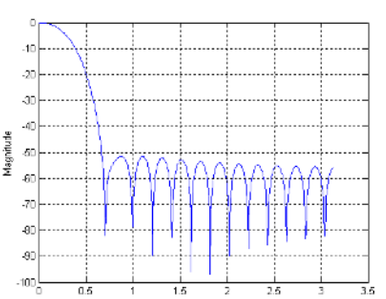


d. Gain in dB of designed IIR filter

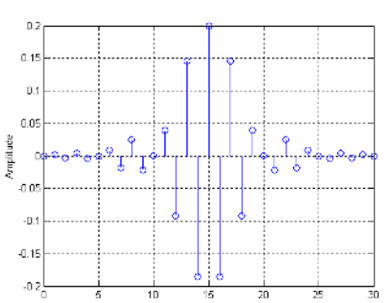
Figure 12. Low-pass signal $x_1(n)$ and high-pass signal $x_2(n)$



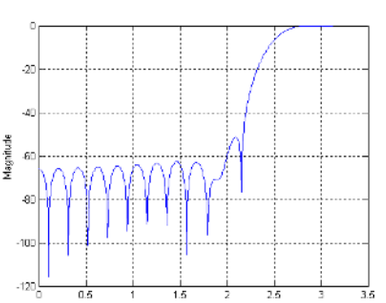
a. Low-pass signal in time-domain.



b. Low-pass signal in frequency domain.

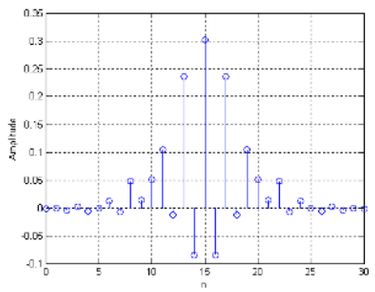


c. High-pass signal in time-domain.

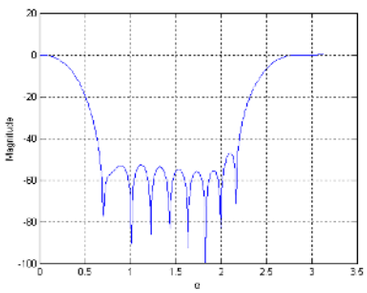


d.. High-pass signal in frequency domain.

Figure 13. Composite signal as the sum of $x_1(n)$ and $x_2(n)$

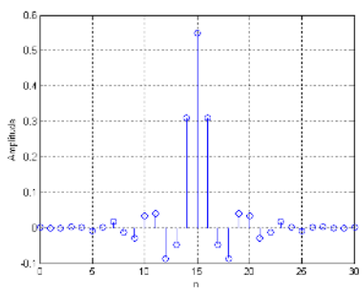


a. *Time–domain signal.*

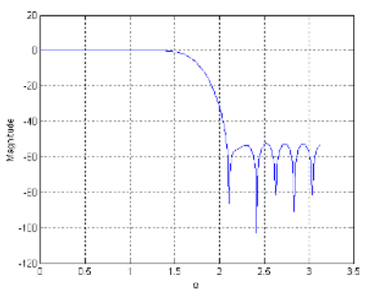


b. Frequency characteristic.

Figure 14. Designed low-pass filter

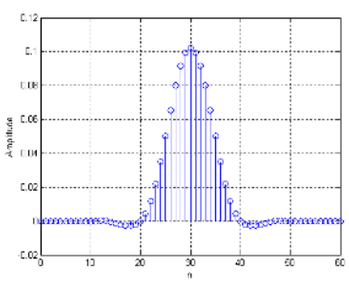


a. Impulse response.

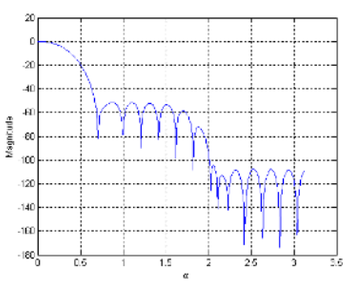


b. Magnitude response.

Figure 15. Filtered signal

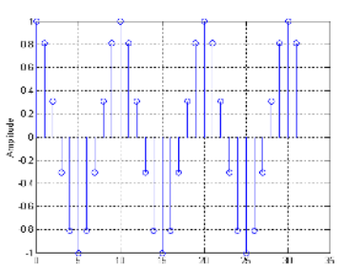


a. Time-domain characteristic.

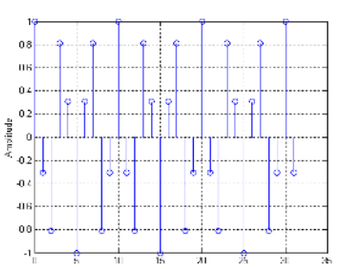


b. Frequency characteristic.

Figure 16. Cosine signals $x_1(n)$ and $x_2(n)$

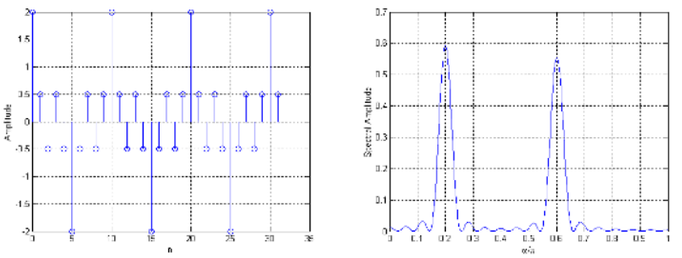


a. Cosine signal $x_1(n)$.



b. Cosine signal $x_2(n)$.

Figure 17. Sum of two cosine signals



a. Time-domain. *b. Frequency domain.*

EXAMPLES OF FILTERING

Example 1

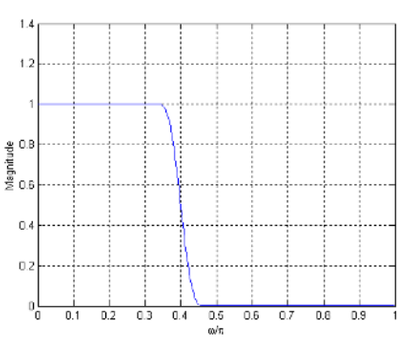
A simple low-pass digital signal $x_1(n)$ is plotted in Figure12.a. Its spectral characteristic is shown in Figure 12b. Time- and frequency-domain representations of a high-pass signal $x_2(n)$ are shown in Figures 12c and 12d.

Suppose we add signals $x_1(n)$ and $x_2(n)$. The result is shown in Figure 13. The composite signal has both low-pass and high-pass components. In order to eliminate the high-pass components we need to pass this signal through a low-pass filter which will preserve only the low-pass components and eliminate the high-pass ones. The low-pass filter is shown in Figure 13. Figure 14 shows the result of filtering of the composite signal.

Example 2

In this example we consider a signal composed of two cosine signals $x_1(n)$ and $x_2(n)$, shown in Figure 16.

Figure 18. Designed low-pass filter



The result of adding these two signals is shown in Figure 17. Two peaks at 0.2π and 0.6π in the spectral characteristic correspond to the cosine components x_1 and x_2, respectively.

Suppose we now apply low-pass filtering to the sum of these two cosine signals. The designed low-pass filter is shown in Figure 18, and the result of filtering is shown in Figure 19. Notice that the second high pass cosine signal has been eliminated.

To eliminate the low-pass cosine signal, we design the high-pass filter shown in Figure 20. The filtered signal is shown in Figure 21.

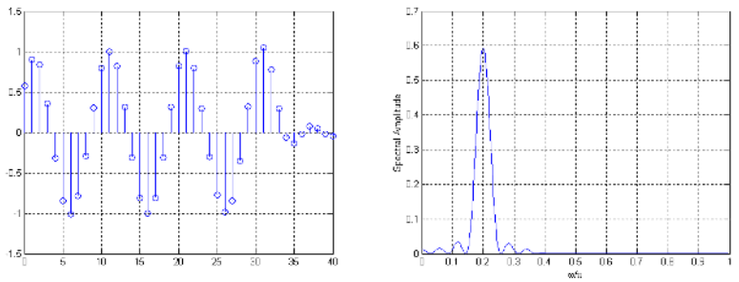
Example 3

The following figure presents an example of a speech signal (McClellan, Schafer, & Yoder, 1998).

We consider one part of the signal (the samples from 1300 to 1500), which is shown in Figure 23a. This figure shows the waveform samples, while Figure 23b. presents the spectral characteristic of this waveform. We apply two low-pass filters. One of them passes all spectral components below 0.25π and eliminates all spectral components higher than 0.3π (Figure 24a). The other filter (Figure 24b) passes only spectral components bellow 0.05π, and attenuates all components higher than 0.1π.

The speech signal filtered by the first filter is shown in Figure 25, while the result of filtering with the second filter is shown in Figure 26. Notice that the resulting signal becomes smoother when higher frequencies are eliminated. By comparing both filtered signals one can notice that even smoother signal can be obtained when lower frequencies are preserved. Therefore, low-pass filtering can be used to remove large fluctuations in the signal.

Figure 19. Filtered signal



a. Time-domain signal. *b. Frequency domain.*

Figure 20. Designed high-pass filter

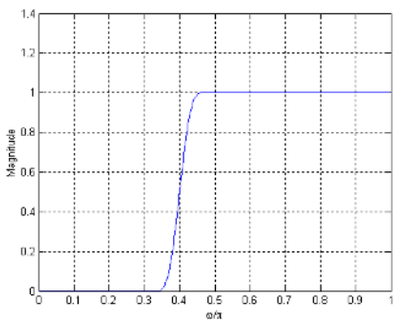
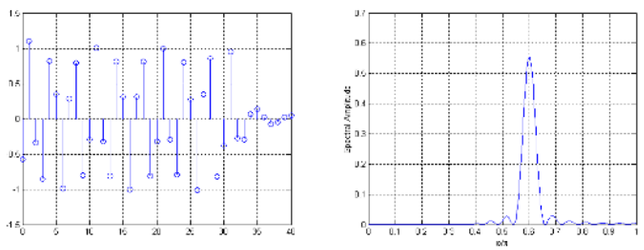


Figure 21. Filtered signal



a. Time domain. *b. Frequency domain.*

Figure 22. Sampled speech waveform

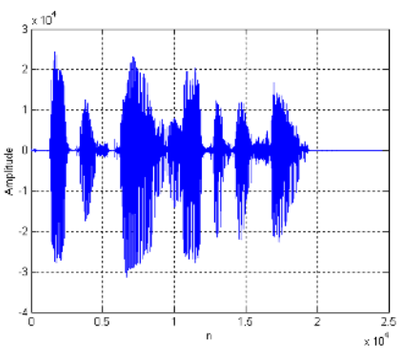
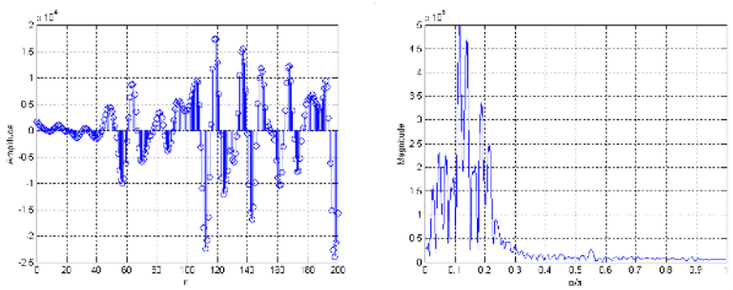
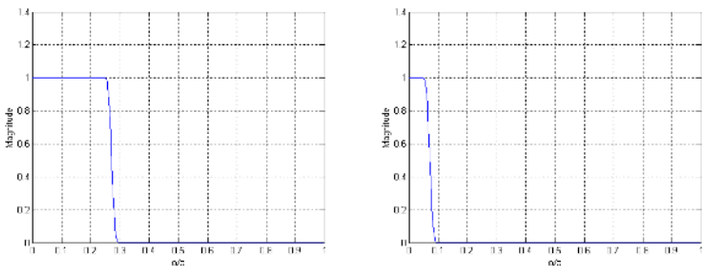


Figure 23. A part of the speech waveform (samples from 1300 to 1500)



a. Time-domain. *b. Frequency domain.*

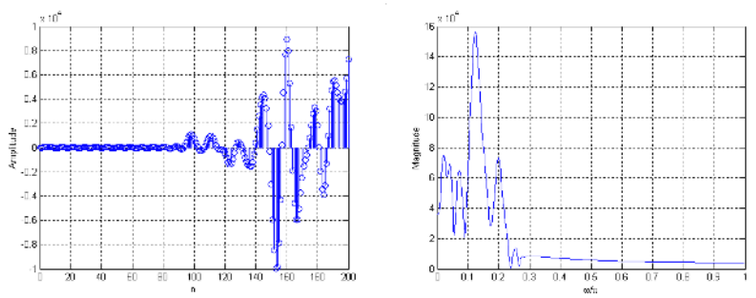
Figure 24. Low-pass filters



a. Magnitude characteristic of the first filter.

b. Magnitude characteristic of the second filter.

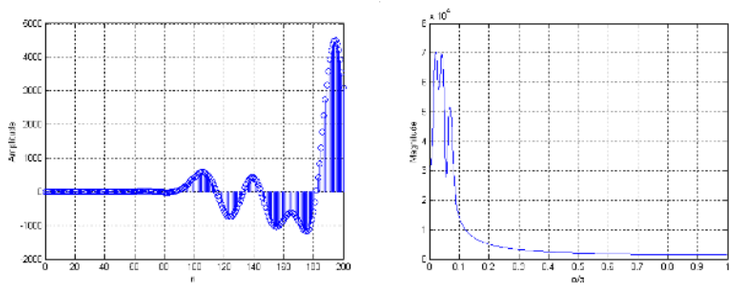
Figure 25. Filtered signal



a. Time domain characteristic.

b . Frequency characteristic.

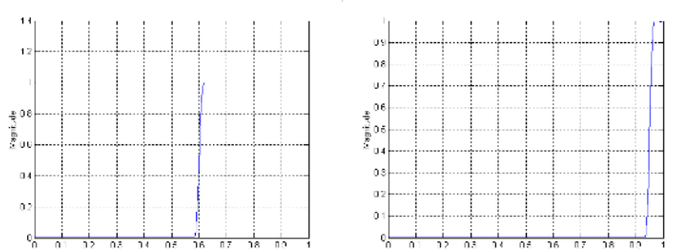
Figure 26. Filtered signal



a. Time domain characteristic.

b . Frequency characteristic.

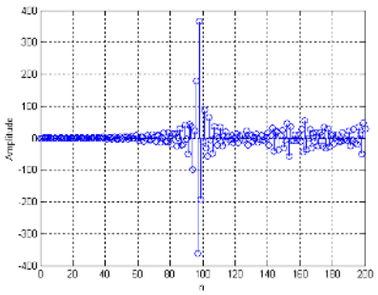
Figure 27. High-pass filters



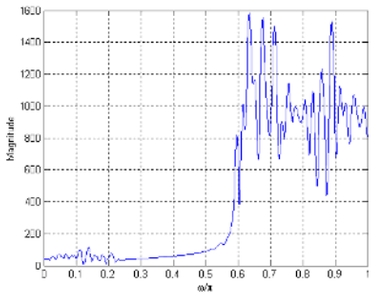
a. Magnitude characteristic. of the first filter.

b. Magnitude characteristic of the second filter.

Figure 28. Filtered signal

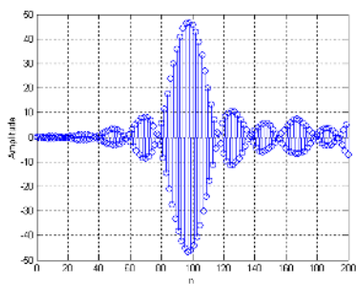


a. Time domain characteristic.

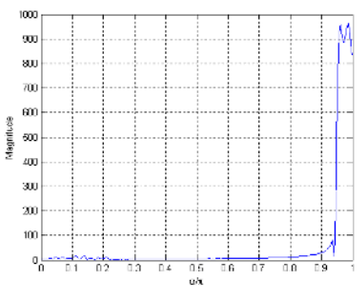


b . Frequency characteristic.

Figure 29. Filtered signal



a. Time domain characteristic.



b . Frequency characteristic.

We now apply two high-pass filters shown in Figure 27 to the speech signal from Figure 23. The result of filtering shown in Figures 28 and 29 demonstrates that the signal in which higher frequency components are preserved is less smooth and has more fluctuations.

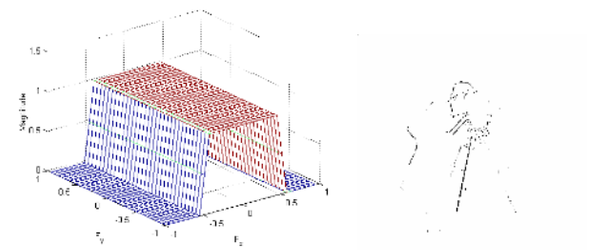
Example 4

In this example we illustrate the effect of filtering of an image. Filtering is used for modifying or enhancing an image, for example to emphasize certain features or remove others. In linear image filtering, two-dimensional filters are used, and they can be obtained from corresponding one-dimensional filters. FIR filters are more convenient for image filtering, because of stability, ease of design and implementation. We apply low pass and high pass filtering to the image signal, (generated in MATLAB), given in Figure 30. The two-dimensional low-pass filter and the result of filtering are shown in Figures 31a and 31b, respectively. The two-dimensional high-pass filter, shown in Figure 30c, is applied to the image, and the resulting image is shown in Figure 30d. We can notice that the effect of low-pass filtering is image smoothing, while high-pass filtering causes enhancement of variations across the image.

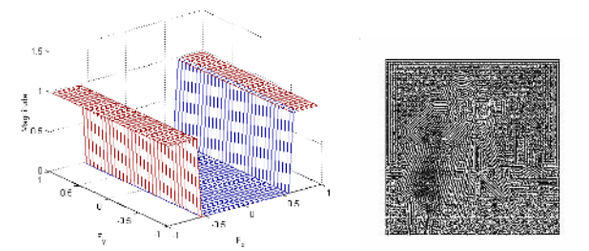
The noise is added to the image and the result is shown in Figure 32a. Two filters are applied to eliminate the noise. Figure 32b shows the result of applying a simple averaging filter, while Figure 32c shows the effect of applying a special filter called

Figure 30. Image signal

Figure 31. Two dimensional low-pass filters and the filtered images



a. Two-dimensional LP filter. *b. Filtered image.*



c. Two-dimensional HP filter. *d. Filtered image.*

Figure 32. Removing the noise from the image

a. The image with added noise.

b. Filtering with averaging filter *c. Filtering with median filter.*

the median filter. Notice that the median filter is much better in removing noise.

CONCLUSION

Digital signal processing lies at the heart of the modern technological development finding the applications in a different areas like image processing, multimedia, audio signal processing, communications, and so on. A system which performs digital signal processing is called a digital filter. The digital filter changes the characteristics of the input digital signal in order to obtain the desired output signal. Digital filters either have a finite impulse response, (FIR), or an infinite impulse response, (IIR). FIR filters are often preferred because of desired characteristics, such as linear phase and no stability problems. The main disadvantage of FIR filters is that they involve a higher degree of computational complexity compared to IIR filters with equivalent magnitude response. In many applications where the linearity of the phase is not required, the IIR filters are preferable because of the lower computational requirements. Over the past several years there have been a number of attempts to reduce the complexity of FIR filters. The design of FIR filters with low complexity and IIR filters with approximately linear phase are the major digital filter design tasks.

REFERENCES

Adams, J.W. & Willson, A.N. (1983). A new approach to FIR digital filter design with fewer multipliers and reduced sensitivity. *IEEE Trans. Circuits and Systems*, *30*, 277-283.

Adams, J.W. & Willson, A.N. (1984). Some efficient digital prefilter structure. *IEEE Trans. Circuits and Systems*, *31*, 260-265.

Bartolo, A., Clymer, B.D., Burgess, R.C., & Turnbull, J.P. (1998). An efficient method of FIR filtering based on impulse response rounding. *IEEE Trans. on Signal Processing*, *46*, 2243-2248.

Coleman, J.O. (2002). Factored computing structures for multiplierless FIR filters using symbol-sequence number systems in linear spaces. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 3*, 3132-3135.

Diniz, P.S.R., da Silva, E.A.B., & Netto, S.L. (2002). *Digital signal processing, system analysis and design.* Cambridge: Cambridge University Press.

Elali, T.S. (2003). *Discrete systems and digital signal processing with MATLAB.* Boca Raton, FL: CRC Press.

Grover, D. & Deller, J.R. (1999). *Digital signal processing and the microcontroller*. NJ: Prentice Hall, Inc.

Ifeachor, E.C. & Jervis, B.E. (2001). *Digital signal processing: A practical approach, second edition.* NJ: Prentice Hall.

Jovanovic-Dolecek, G. & Mitra, S.K. (2002). Design of FIR lowpass filters using stepped triangular approximation. *The 5th Nordic Signal Processing Symposium, NORSIG 2002,* on board Hurtigruten M/S Trollfjord, Norway.

Kuc, R. (1988). *Introduction to digital signal processing,* New York: McGraw-Hill.

Liu, M.C., Chen, C.L., Shin, D.Y., Lin, C.H., & Jou S.J. (2001). Low-power multiplierless FIR filter synthesizer based on CSD code. *The 2001 International Symposium on Circuits and Systems, ISCAS 2001*, 4, 666-669.

The MathWorks Incorporation (1997). *Image processing toolbox user's guide.* Natick: The MathWorks.

McClellan, J.H., Schafer, R.W., & Yoder, M.A. (1998). *DSP First: A multimedia approach.* NJ: Prentice-Hall.

Mitra, S.K. (2001). *Digital signal processing: A computer-based approach, second edition.* New York: The McGraw-Hill Companies.

Oppenheim, A.V. & Schafer, R.W. (1999). *Discrete-time signal processing* (second edition). NJ: Prentice-Hall.

Proakis, J.G. & Manolakis, D.G. (1996). *Digital signal processing: Principles, algorithms and applications.* Englewood Cliffs, NJ: Prentice-Hall.

Ramakrishan, K.V & Gopinathan, E. (1989). Design of FIR filters with reduced computations. *Circuits Systems Signal Proc*., *8*, 17-23.

Silva, J.M. & Jovanovic-Dolecek, G. (1999). Discrete-time filters. In J.G. Webster (Ed.), *Wiley encyclopedia of electrical and electronics engineering,* (Vol. 5, pp. 631-643). New York: John Wiley & Sons.

Smith, D. (2001). *Digital signal processing technology: Essentials of the communications revolution*. New York: Amer. Radio Relay League.

Smith, S. (2002). *Digital signal processing: A practical guide for engineers and scientists*. Newnes.

Stearns S.D. (2002). *Digital signal processing with examples in MATLAB*. Boca Raton, FL: CRC Press.

Stein, J. (2000). *Digital signal processing: A computer science perspective*. New York: Wiley-Interscience.

Tai, Y.L. & Lin, T.P. (1992). Design of multiplierless FIR filters by multiple use of the same filter. *Electronics Letters*, *28*, 122-123.

Yli-Kaakinen, J. & Saramaki, T. (2001). A systematic algorithm for the design of multiplierless FIR filters. *The 2001 International Symposium on Circuits and Systems, ISCAS 2001*, *2*, 185-188.

White, S. (2000). *Digital signal processing: A filtering approach*. Delmar Learning.

KEY TERMS

Digital Filter: The digital system which performs digital signal processing i.e., transforms an input sequence into a desired output sequence.

Digital Signal: A discrete-time signal whose amplitude is also discrete. It is defined as a function of an independent, integer-valued variable n. Consequently, a digital signal represents a sequence of discrete values, (some of which can be zeros), for each value of integer n.

Digital Signal Processing: Extracts useful information carried by the digital signals and is concerned with the mathematical representation of the digital signals and algorithmic operations carried out on the signal to extract the information.

FIR Filter: A digital filter with a finite impulse response. FIR filters are always stable. FIR filters have only zeros (all poles are at the origin).

IIR Filter: A digital filter with an infinite impulse response. IIR filters always have poles and are stable if all poles are inside the unit circle.

Impulse Response: The time domain characteristic of a filter and represents the output of the unit sample input sequence.

Magnitude Response: The absolute value of the Fourier transform of the unit sample response. For a real impulse response digital filter, the magnitude response is a real even function of the frequency.

Phase Response: The phase of the Fourier transform of the unit sample response. For a real impulse response digital filter, the phase response is an odd function of the frequency.

Signal: Any physical quantity that varies with changes of one or more independent variables which can be any physical value, such as time, distance, position, temperature, and pressure.

Stable Filter: A filter for which a bounded input always results in a bounded output.

Digital Video Broadcasting (DVB) Applications

Ioannis P. Chochliouros
Hellenic Telecommunications Organization S.A. (OTE), Greece

Anastasia S. Spiliopoulou-Chochliourou
Hellenic Telecommunications Organization S.A. (OTE), Greece

George K. Lalopoulos
Hellenic Telecommunications Organization S.A. (OTE), Greece

INTRODUCTION

The topic of Digital Video Broadcasting (DVB) applications (including both infrastructures and services) is a very broad one. It encompasses not only the transmission and distribution of television-program material in digital format over various media, but also a range of related features designed to exploit the capabilities of all possible underlying technologies. Within a fully converged environment, DVB can contribute to the effective penetration and adoption of a variety of enhanced multimedia services (Fenger & Elwood-Smith, 2000) based on various forms and types of content (with major emphasis given to the audiovisual sector; European Commission, 1999). Moreover, DVB intends to support optimized solutions for different communications platforms. Europe has adopted DVB for use across all relevant technical platforms. In fact, Europe has the highest density of TV homes in the world and is leading the deployment of digital TV (European Commission, 2003a) through DVB. The focus provided by a common set of technical standards and specifications throughout the European Union (EU) has given a market advantage and spurred the deployment of digital television services. Market expansion will be determined by the rate at which broadcasters are enabled to develop services and by the cost of Set-Top Boxes (STBs) or integrated television equipment.

Current European policies have provided a market advantage and accelerated the development and deployment of modern applications based upon specific features of existing infrastructures, also taking into account the needs of the European citizens with those of the media, telecommunications, and equipment industries (European Commission, 2002).

In particular, as for the European framework, all applicable technical specifications for digital broadcasting are currently promoted under the scope of the DVB Project. Those specifications are then offered for standardization to the relevant standards body, that is, ETSI (European Telecommunications Standards Institute) and/or CENELEC (European Committee for Electrotechnical Standardization): The latter deals with the consumer equipment aspects while the former with all other aspects.

CURRENT STANDARDIZATION INITIATIVES: THE DVB PROJECT

The DVB Project (officially formed in September 1993) is a market-led consortium of public- and private-sector organizations in the television industry, comprising over 300 broadcasters, manufacturers, network operators, software developers, and regulatory bodies from more than 35 countries worldwide who are committed to designing global technical standards for the delivery of digital television. Its aim is to establish the framework for the introduction of MPEG-2- (Moving Pictures Experts Group-2) based digital television services. All promoted works foster market-led systems, which meet the real needs and economic circumstances of the consumer electronics and broadcast industry (Reimers, 2000).

In the course of recent years, a considerable list of specifications has been developed very success-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

fully; these can be used for broadcasting all kinds of data as well as sound, accompanied by possible types of auxiliary information. Some of the specifications aim at the installation of appropriate bidirectional communication channels via the exploitation of existing networks (Nera Broadband Satellite, 2002). Due to the huge complexity of the surrounding "environment", different factors have to be taken into account when planning services or equipment. However, the aim is the creation of a coordinated digital broadcast market for all service delivery media. The Project is not a regulator or "government-driven" (top-down) initiative. Working with tight timescales and strict market requirements, the project intends to achieve considerable economy of scale, which in turn ensures that, toward the expected transformation of the industry to digital technologies, broadcasters, manufacturers, and, ultimately, the viewing public will benefit.

The work does not intend to specify an interaction channel solution associated to each broadcast system, especially because the interoperability of different delivery media is desirable. Therefore, any potential solution for the interaction channel applies to satellite DVB (DVB-S), Cable DVB (DVB-C), Terrestrial DVB (DVB-T), Master Antenna Television (MATV), Satellite Master Antenna Television (SMATV), Microwave or MMDS (Multi-Channel Multipoint Distribution Systems) DVB (DVB-MS/MC), or any future DVB broadcasting or distribution system.

As a consequence, progress realized up to now has developed a complete family of interrelated television systems for all possible transmission media at all quality levels (from standard definition to high definition, including the enhanced definition 16/9 format currently being widely deployed in Europe). The standards also cover a range of tools (Valkenburg & Middleton, 2001) for added-value services such as pay-per-view, interactive TV (i-TV), data broadcasting, and high-speed, "always-on" Internet access.

OPTIMIZED SOLUTIONS FOR DIFFERENT TECHNICAL PLATFORMS

The basic components of DVB are the use of MPEG-2 packets as digital "data containers" and the critical, relevant Service Information (SI) surrounding and identifying those packets. DVB can deliver to the home almost anything that can be digitized, whether this is High-Definition TV (HDTV), multiple-channel Standard-Definition TV (SDTV, i.e., PAL, NTSC, or SECAM), or broadband multimedia-data and interactive electronic-communications services.

The video, audio, and other data are inserted into fixed-length MPEG Transport Stream (TS) packets. Packetized data constitutes the payload, which can carry any combination of MPEG-2 (video and audio).

Thus, service providers are free to deliver anything from multiple-channel SDTV, 16:9 wide-screen Enhanced-Definition Television (EDTV), or single channel HDTV to multimedia data broadcast-network services and Internet over the air.

The complete "system" can be seen as a "functional block" of clusters of equipment performing the adaptation of the baseband signals, from the output of the MPEG-2 transport multiplexer to the corresponding channel characteristics. The following processes are generally applied to the data stream: (a) transport multiplex adaptation and randomization for energy dispersal; (b) outer coding (e.g., Reed-Solomon); (c) convolutional interleaving; (d) inner coding (e.g., punctured convolutional coding); (e) baseband shaping for modulation; and (f) modulation.

However, to make a fair assessment of the impact of DVB, it is essential to consider its presence on three fundamental and distinct platforms, that is, satellite, cable, and terrestrial or microwave.

The satellite system, DVB-S, is the oldest and most established of the DVB standards family, and it arguably forms the "core" of the great success in the market. The satellite system is designed to cope with the full range of satellite transponder bandwidths and services (ETSI, 2003a). The DVB-C cable system (ETSI, 1998) is based on DVB-S, but the modulation scheme used is Quadrature Amplitude Modulation (QAM) instead of quadrature phase-shift keying (QPSK; as in the previous case). The system is centered on 64-QAM, but it also allows for lower and higher level systems. In each case, there is a trade-off between data capacity and robustness of data.

Under a similar approach, the terrestrial DVB-T system specification is based on MPEG-2 sound and

vision coding (ETSI, 2004a). The modulation system combines OFDM (Orthogonal Frequency Division Multiplexing) with QPSK and QAM. OFDM uses a large number of carriers, which spread the information content of the signal. Used very successfully in DAB (Digital Audio Broadcasting), OFDM's major advantage is that it thrives in a very strong multipath environment (ETSI, 2002), making it possible to operate an overlapping network of transmitting stations with a single frequency, especially in mobile reception conditions.

The DVB Multipoint Distribution System uses microwave frequencies for direct distribution to viewers' homes (ETSI, 1999). Its first version, DVB-MC, is based on the DVB-C cable delivery system, and will therefore enable a common receiver to be used for both cable transmissions and this type of microwave transmission. DVB-MC makes use of frequencies below 10 GHz. Its second version, DVB-MS, is based on the DVB-S satellite delivery system. DVB-MS signals can therefore be received by DVB-S satellite receivers, which need to be equipped with a small MMDS frequency converter rather than a satellite dish. DVB-MS makes use of frequencies above 10 GHz.

Community antenna systems are important in many markets. DVB-CS is the DVB digital SMATV system (ETSI, 1997), adapted from DVB-C and DVB-S. The primary consideration of such a system is the transparency of the SMATV head-end to the digital TV multiplex from a satellite reception without baseband interfacing, delivering the signal to the user's Integrated Receiver Decoder (IRD; typically the set-top box). In general, technology can permit the establishment of a simple and cost-effective head-end for the consumer.

Data broadcasting (ETSI, 2004b) is designed to allow operators to download software and applications over satellite, cable, or terrestrial links, for example, to deliver Internet services over broadcast channels (using IP [Internet Protocol] tunneling) and to provide interactive TV. MPEG-2 DSM-CC (Digital Storage Media-Command and Control) has been chosen as the core of the related specification. The result is based on a series of four profiles, each one serving a specific application area.

These are listed as follows.

1. **Data Piping:** This is a simple, asynchronous, end-to-end delivery of data through DVB-compliant broadcasting networks.
2. **Data Streaming:** This profile can support services requiring a streaming-oriented, end-to-end delivery of data in either an asynchronous, synchronous, or synchronized way.
3. **Multiprotocol Encapsulation:** This explicit profile supports services requiring the transmission of datagrams of communication protocols.
4. **Data Carousels:** They support services requiring the periodic transmission of data modules.

GENERAL FEATURES OF DVB SYSTEMS

The fundamental features of DVB systems can be considered as: openness, interoperability, flexibility, market-led nature, and innovative nature.

Openness

DVB systems are developed through consensus in the standardization working groups to implement innovative features conformant to user requirements. Once standards have been published through the procedures of ETSI, these are available at a nominal cost for anyone at the global level. In fact, open standards provide the manufacturers an opportunity to freely implement innovative and value-added services independently of the kind of the underlying technology (Reimers, 2000).

Interoperability

Because the reference standards are open, all manufacturers deploying compliant systems are able to guarantee that their equipment will interwork with other similar equipment. As standards are designed with a maximum amount of commonality and based on the common MPEG-2 coding system, they may be effortlessly carried from one medium to another to minimize development and receiver costs; in particular, such a perspective provides a significant advantage as it offers opportunities for simple, transparent, and effective signal distribution

in various technical platforms. DVB signals can move easily and inexpensively from one transmission and reception means to another with minimum processing, thus promoting convergence and technological neutrality.

Interfacing

Interfacing is the key to interoperability (European Commission, 2003a), and DVB has established a detailed set of professional and consumer receiver-interface specifications to ensure that head-end equipment can originate from different sources and can be used without limitations in the markets, thus promoting options for competition and growth.

The related specifications include interfaces to Plesiochronous Digital Hierarchy (PDH) and Synchronous Digital Hierarchy (SDH) networks, as well as for CATV (Community Antenna Television) and SMATV head-ends and similar professional equipment. Such interfaces support compatibility options and provide assurance that consumer equipment can be adequately connected to future in-home digital networks.

Flexibility

The use of MPEG-2 packets as "data containers" (ETSI, 2004b; also considering the relevant SI for their identification) provides major benefits: DVB can deliver to the home almost all forms of digital information (i.e., from "traditional" TV programs to multimedia and interactive services). The option for flexibility may take into account existing differences in priorities between operators regarding capacity (e.g., number of channels and quality) and coverage, as well as probable differences in receiving conditions. It is evident that flexibility can affect very strongly the market in multiple sectors (technical, commercial, financial, business, regulatory, etc.; Norcontel Ltd., 1997).

Market-Led Nature

Contrary to earlier, similar initiatives in Europe and the United States, works carried out within the wider DVB context intend to meet certain well-defined needs and some prescribed requirements as imposed by the market itself. Moreover, the fact of the "active" participation of a variety of market players originating from various sectors (e.g., industry manufacturers, network operators, service providers, broadcasters, etc.) can be a prerequisite to guarantee that the proposed solutions will be fully developed to satisfy both current market requirements and requests. In such a way, both businesses and citizens can have access to an inexpensive, world-class communications infrastructure and a wide range of (multimedia) services (UK's Consumers' Association, 2001).

Innovative Nature

Digital technology and the convergence of various media are going to introduce many more alternatives and special facilities besides the traditional one-to-many form of communication that we understand by "television" today. The convergence between telecommunications, broadcasting, and information technologies affects very drastically (European Commission, 2002, 2003a) the corresponding transmission and introduces revolutionary options for the markets. New services making use of the advanced features of digital (television) technology will present many-to-one, many-to-many, and one-to-one communication. In combination with an interactive return channel (using an interface to a mobile phone or a Personal Digital Assistant [PDA], for example), digital receivers will be able to offer users a variety of enhanced services, from simple interactive quiz shows to Internet over the air and a mix of television and Web-type content. High-quality mobile television reception (unachievable with existing analogue systems) is expected to be a reality in a short-term perspective.

Interactivity

Many of the service offers possible in the DVB world will require some form of enhanced interaction (Neale, Green, & Landovskis, 2001) between, for example, the end user and either the program provider or the network operator. This sort of "interaction" may consist of the transmission of just a few commands, but may be extremely extensive and thus resemble communication via the Internet (especially

to distribute content and modern applications from electronic communications, such as broadband e-learning, e-health, e-entertainment, etc.). As innovation evolves, interactive TV has been identified as one of the key areas ideally suited to an entirely digital transmission system. DVB has developed comprehensive plans for such an introduction: The result is a set of specifications for interactive services and a series of network-specific specifications designed to suit the needs of the physical characteristics of the individual media (ETSI, 1997, 1998, 1999, 2002, 2003a, 2004a, 2004b).

Other Aspects

In order to realize a meaningful approach for development, a number of parameters must also be considered. These may include, among others, the following.

1. More enhanced safety and security options (especially if referred to activities such as electronic commerce or electronic transactions). Security can also cover other requirements to avoid damage to people or to installations in the surrounding area.
2. Adequate RF-(Radio Frequency) performance options to provide a certain degree of Quality of Service (QoS). In the same context, particular emphasis will be given to avoiding any noise and/or interference effects probably produced by the surrounding environment.
3. Control and Monitoring Functions (CMFs) to guarantee efficient overview of the entire system "entity" and of the services offered.

As work progresses continuously through the introduction of innovative applications, the next phase of DVB emphasizes upon the impact of digital television and the convergence effect in the home. The Multimedia Home Platform (MHP) is a specific example of such a revolutionary innovation (ETSI, 2003b) aiming to standardize the main software and hardware interfaces in the home. Based around a sort of Java Application Programming Interface (API), it provides an open environment for enhanced applications and services. The rewards for the industry are expected to be enormous and high. MHP is a tremendous opportunity to facilitate the passage from today's vertical markets, using proprietary technologies, toward horizontal markets based on open standards to benefit consumers and market players (Digital Video Broadcasting, 2001). At the political level, the European Commission (2003b) has undertaken support for MHP implementation and results are expected to be encouraging in the near future.

CONCLUSION: THE WAY FORWARD

DVB applications generally offer tremendous possibilities for (a) technical solutions suitable to respond to current and/or forthcoming commercial requirements, including a wide range of service options ranging from LDTV (Low Definition Television) to HDTV; (b) improved transmission and reception quality; (c) the flexibility to reconfigure the available data capacity between different service options (exchange between quality and quantity with related cost consequences); (d) innovative, interactive broadband services originating from the wider information society sector; and (e) HDTV services when cost-effective and convenient equipment become available for the consumer.

To achieve the new potential benefits, positive steps should be taken to facilitate consumer choices as well as the migration to new applications. Such steps could include, among others: (a) attractive program offers plus new services aimed at a suitably defined audience (including specific service features that other transmission forms will not be able to offer); (b) the convenient introduction of appropriate digital receivers (also including set-top boxes) at appropriate prices; (c) the forwarding of new technology implementations able to satisfy new applications; and (d) the establishment and/or update of suitable European DVB-based standards to enforce the currently offered technical solutions.

The DVB option offers multiple technical aspects to specify a broadcasting and/or distribution system (as well as the corresponding interaction-channel solution) that should be suitable to promote any relevant applications. Within the DVB context, there is the possibility to consider various alternatives for different transmission media at all quality

levels to cover a range of interactive services, data broadcasting, and so forth.

Simultaneously, user requirements outline market parameters for the selected system (i.e., price band, user functions, etc.), and they are used as guidelines for the technical specification process in order to maintain a practical perspective. The next step to be performed in the market, especially via the MHP, is expected to open new horizons for investment and growth.

REFERENCES

Digital Video Broadcasting. (2001). *DVB commercial module-multimedia home platform: User and market requirements* (DVB Blue Book A062). Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI).

ETSI. (1997). *TR 101 201 V1.1.1-DVB. Interaction channel for Satellite Master Antenna TV (SMATV) distribution systems: Guidelines for versions based on satellite and coaxial sections*. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI).

ETSI. (1998). *ETS 300 800-DVB: DVB interaction channel for cable TV distribution systems*. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI).

ETSI. (1999). *EN 301 199 V1.2.1: Interaction channel for Local Multi-Point Distribution Systems (LMDS)*. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI).

ETSI. (2002). *EN 301 958 V1.1.1-DVB: Interaction channel for digital terrestrial television (RCT) incorporating multiple access OFDM*. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI).

ETSI. (2003a). *EN 301 790 V1.3.1-DVB: Interaction channel for satellite distribution systems*. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI).

ETSI. (2003b). *TS 201 812 V1.1.1-DVB: Multimedia Home Platform (MHP) specification 1.0.3*. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI).

ETSI. (2004a). *EN 300 744 V1.5.1-DVB: Framing structure, channel coding and modulation for digital terrestrial television*. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI).

ETSI. (2004b). *EN 301 192 V1.4.1-DVB: DVB specification for data broadcasting*. Sophia Antipolis, France: European Telecommunications Standards Institute (ETSI).

European Commission. (1999). *Communication on the development of the market for digital television in the European Union [COM (1999) 540 final, 09.11.1999].* Brussels, Belgium: European Commission.

European Commission. (2002). *Communication on eEurope 2005: An information society for all [COM(2002) 263 final, 28.05.2002].* Brussels, Belgium: European Commission.

European Commission. (2003a). *Communication on barriers to widespread access to new services and applications of the information society through open platforms in digital television and third generation mobile communication [COM (2003) 410 final, 09.07.2003].* Brussels, Belgium: European Commission.

European Commission. (2003b). *Communication on the transition from analogue to digital broadcasting (from digital "switchover" to analogue "switch-off") [COM (2003) 541 final, 17.09.2003].* Brussels, Belgium: European Commission.

Fenger, C., & Elwood-Smith, M. (2000). *The fantastic broadband multimedia system.* Geneva, Switzerland: The Fantastic Corporation.

Neale, J., Green, R., & Landovskis, L. (2001). Interactive channel for multimedia satellite networks. *IEEE Communications Magazine, 39*(3), 192-198.

Nera Broadband Satellite (NBS) A.S. (2002). *Digital video broadcasting return channel via satellite (DVB-RCS) — Background book.* Oslo, Norway: NBS.

Norcontel Ltd. (1997). *Economic implications of new communication technologies on the audio visual markets* (Final Report). Brussels, Belgium: NERA, Screen Digest, Stanbrook & Hooper.

Reimers, U. (2000). *Digital Video Broadcasting: The international standard for digital television.* Berlin and Heidelberg, Germany; New York: Springer-Verlag.

UK's Consumers' Association. (2001). *Turn on, tune in, switched off: Consumers' attitudes to digital TV.* London: UK's Consumers' Association.

Valkenburg, M. E. Van, & Middleton, W. M. (2001). *Reference data for engineers: Radio, electronics, computer, and communications.* Boston: Newnes.

KEY TERMS

Baseband: (1) In radio communications systems, the range of frequencies, starting at 0 Hz (Direct Current - DC) and extending up to an upper frequency as required to carry information in electronic form, such as a bit stream, before it is modulated onto a carrier in transmission or after it is demodulated from a carrier in reception. (2) In cable communications, such as those of a local-area network (LAN), a method whereby signals are transmitted without prior frequency conversion.

Carrier: A transmitted signal that can carry information, usually in the form of modulation.

Digital Television (DTV): The term adopted by the FCC (Federal Communications Commission - USA Regulatory Body) to describe its specification for the next generation of broadcast-television transmissions. DTV encompasses both HDTV and STV.

Digital Video Broadcasting (DVB): It originally meant television broadcasting using digital signals (as opposed to analogue signals), but now refers to broadcasting all kinds of data as well as sound, often accompanied by auxiliary information and including bidirectional communications.

European Telecommunications Standards Institute (ETSI): An organization promulgating engineering standards for telecommunications equipment. The secretariat is at Valbonne, France (http://www.etsi.org/).

High-Definition Television (HDTV): A new type of television that provides much better resolution than current televisions based on the NTSC standard. There is a number of competing HDTV standards, which is one reason why the new technology has not been widely implemented. All of the standards support a wider screen than NTSC and roughly twice the resolution. To pump this additional data through the narrow TV channels, images are digitalized and then compressed before they are transmitted and then decompressed when they reach the TV. HDTV can offer bit rates within the range of 20 to 30 Mbit/s.

Interaction Channel (IC): A bidirectional channel established between the service provider and the user for interaction purposes.

Low Definition Television (LDTV): A type of television providing a quality of image usually compared to VHS (Video Home System); this practically corresponds to the collection of television fragments videotaped directly from the TV screen. The bit rate offered is 1.5 Mbit/s (1.15 Mbit/s for the video only), which corresponds to the bit rate offered by the original standard MPEG-1.

MPEG-2: Refers to the standard ISO (International Organization for Standardization) / IEC (International Electrotechnical Commission) 13818. (Systems coding is defined in part 1 of the standard. Video coding is defined in part 2 of the standard. Audio coding is defined in part 3 of the standard.)

Multimedia Home Platform (MHP): A DVB project to devise specifications for a home-network architecture and a next-generation, open set-top box using a standardized interactive application program interface (http://www.mhp.org/).

Quadrature Amplitude Modulation (QAM): A method of modulating digital signals onto a radio-frequency carrier signal involving both amplitude and phase coding.

Quadrature Phase-Shift Keying (QPSK): A method of modulating digital signals onto a radio-frequency carrier signal using four phase states to code two digital bits.

Digital Watermarking Based on Neural Network Technology for Grayscale Images

Jeanne Chen
HungKuang University, Taiwan

Tung-Shou Chen
National Taichung Institute of Technology, Taiwan

Keh-Jian Ma
National Taichung Institute of Technology, Taiwan

Pin-Hsin Wang
National Taichung Institute of Technology, Taiwan

BACKGROUND: WATERMARKING

Great advancements made on information and network technologies have brought on much activity on the Internet. Traditional methods of trading and communication are so revolutionized that everything is quasi-online. Amidst the rush to be online emerge the urgent need to protect the massive volumes of data passing through the Internet daily. A highly dependable and secure Internet environment is therefore of utmost importance.

A lot of research has been done in which watermarking has become an important field of research for protecting data. Watermarking is a technique to hide or embed watermark data in a host data (Cox & Miller, 2001; Martin & Kutter, 2001). The embedded watermark could be retrieved at an appropriate time to be used as proof for rightful ownership when one is in question. Both the watermark and the host data could be any media format such as document, still-image, video, audio, and more. The main concentration for this article is on the still-image. The criteria for watermarking technique are imperceptibility and robustness (Cox, Miller, & Bloom, 2000; Hwang, Chang, & Hwang, 2000; Silva & Mayer, 2003). The embedded watermark must not be easily detectable (imperceptible) so as to discourage hacking; once detected, it must not be easy to decrypt. In some instances where hacking are acts with malicious intents—the watermark must withstand (robustness) these attacks and other attacks such as normal image manipulations like sizing, rotations, cropping, and more (Du, Lee, Lee, & Suh, 2002; Lin, Wu, Bloom, Cox, Miller, & Lui, 2001). Robustness here implies that the watermark can still be recovered after suffering attacks of sorts (Miller, Doerr, & Cox, 2004; Niu, Lu, & Sun, 2000; Silva & Mayer, 2003).

The host image can be manipulated for watermark embedding; either in spatial or frequency domain (Gonzalez & Wood, 2002). In spatial domain, an image is perceived as is—but is digitally represented in terms of pixels. Each pixel reflects a spectrum of colors that is perceptible by the human eye system (HVS). In frequency domain, the image is confined to high, medium, and low frequencies with the HVS being less sensitive to high frequency and more sensitive to low frequency. The proposed watermark technique for this article is interested in embedding a watermark in the frequency domain. In the frequency domain, the embedded watermark is less vulnerable to attacks; be it intentional or unintentional. Therefore, the image will be transformed to its frequency domain using the discrete cosine transformation (DCT) (Liu et al., 2002; Hwang et al., 2000).

Furthermore, we are also interested in applying the neural network technology to embed and extract a watermark. By applying the neural network to embed the watermark, we hope to disperse the watermark such that it can be more securely hidden, not easy to decrypt and imperceptible. Neural net-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

work will again be applied to extract the embedded watermark bits. The train and retrain characteristic could be used to increase the amount of extracted watermark; thereby increasing the chances of getting better quality extracted watermark.

NEURAL NETWORK (NN) TO ENHANCE WATERMARKING

Some of the most popular neural networks (NN) include the Back-propagation Network (BPN), the Probabilistic Neural Network (PNN) and the Counter Propagation Network (CPN). Although the different NNs are devoted to different applications, the basic fundamentals remain in all applications as to how to best apply NN's dynamic learning and error tolerant capacity to get the most accurate results (Davis & Najarian, 2001; Zhang, Wang, & Xiong, 2002). For this article, we are only interested in BPN. BPN is unique for its train and self-train characteristic which can produce more precise trained values. This special characteristic is useful for improving on the watermarking technique such as secure embedment (Hwang et al., 2000) or enhancing the extracted watermark (Tsai, Cheng, & Yu, 2000).

Using NN to Enhance Watermark Embedment

In Hwang et al.'s (2000) method, the watermark will be embedded in the frequency domain. As shown in Figure 1, the host image will first be divided into blocks. These blocks will be individually discrete cosine transformed (DCT) to their frequency domains. The blocks will be scanned in zigzag order to be input as variables to BPN to train for anticipated outputs. Figure 2 shows an example with inputs, {AC1, AC2, …, AC9} with anticipated output {AC12}. The anticipated outputs will be the weighted values used to retrain for a new set of outputs. The final outputs will be paired with the watermark bits {0, 1} to complete the embedding process. Finally, the image will undergo inverse DCT (IDCT) back to spatial domain.

For extraction, the same process will be repeated to divide the image into blocks. Each block will undergo DCT to the frequency domain. The weighted values recorded during BPN training will be used to scan the blocks in zigzag order for the watermark bits. No NN will be applied to the extracted watermark. After watermark had been extracted, the image will be inverse DCT back to spatial domain.

Using NN to enhance the extracted watermark.

In Tsai et al.'s (2000) proposed method the watermark will be first translated into a grouping of 0 and 1 bits. As bits from the grouping are being embedded, they will be tagged. As shown in Figure 3, a 32×32 black and white watermark was embedded into the blue pixels of a 480×512 color host image in spatial domain. The watermark was first converted into bits group *S* which will be randomly encrypted as

Figure 1. Dividing the host image into blocks for DCT

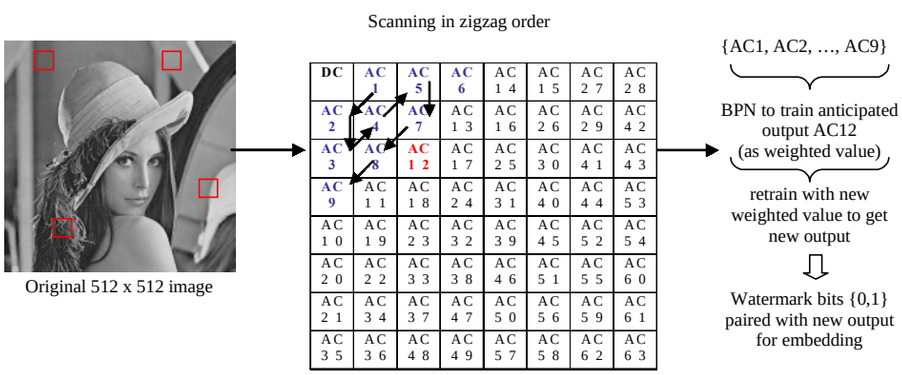
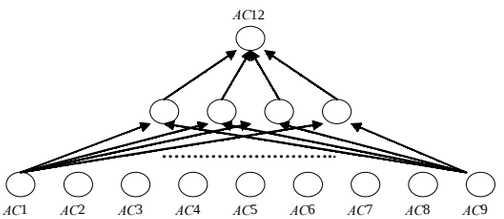

Scanning in zigzag order
Original 512 x 512 image
{AC1, AC2, …, AC9}
BPN to train anticipated output AC12 (as weighted value)
retrain with new weighted value to get new output
Watermark bits {0,1} paired with new output for embedding

Figure 2. Diagram for BPN training (Hwang et al., 2000)



they were being embedded and tagged by the *H* bits. *H* will be embedded together with the watermark bits.

For the extraction process, the embedded *H* will be used to identify the tag locations on the watermark. Details on *H* are available only to the authorized users. NN will be applied to extract a more accurate *H* and then, *S* to get a better quality in the extracted watermark. The same identical random configuration will be used to locate the embedded data *H*. Once an embedded data had been located, its adjacent grids (Figure 4) together with the embedded data will be used to train a network model. Once the NN training is completed, the process is repeated for the embedded watermark data *S*. Outputs from NN will decide on embedding 0's or 1's.

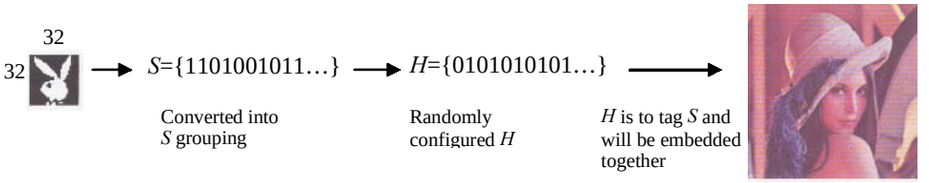
PROPOSED WATERMARKING TECHNIQUE

The proposed watermarking technique will embed watermark in the frequency domain and encrypt hiding with BPN as in Hwang et al. (2000), and enhance the extracted watermark with BPN as in Tsai et al. (2000). By combining the best ideas from both groups of researchers, we have a robust and imperceptible watermark algorithm.

As seen in Figure 5, a grayscale image will be embedded with a 44×44 grayscale watermark (Figure 6). The watermark is first divided into units of 4×4 blocks (Figure 7). Each block will be DCT transformed from spatial to frequency domain. A random sequencer is used to disperse the watermark before embedding. The sequencing rule makes it difficult for any hacker to tamper with the embedded data even if its location is known. Once dispersing is completed, the coefficients from DCT will become disarrayed. The coefficients will be grouped together such that; coefficients greater than 320 will be grouped to 320, and those below -310 to -310. Those falling within the range of -310 and 320 will be divided into groups of 10 such that; coefficients within 5 and 15 to 10, 15 and 25 to 20, and so on. The grouping set will be a subset of {-310, -300, …, 10, 20, …, 310, 320}. Next the elements will be paired such that; –310 with 0, -300 with 1, and so on until 320 to 63. The pairing numbers will be converted into a six-bit binary expression for 0 = {000000}, 1 = {000001}, and so on. {000000} will be embedded instead of -310, and {111111} instead of 320.

Once coding for the watermark is completed, embedding into the original 1024×1024 grayscale image begins. Similar to Tsai et al.'s (2000) *H*2 grouping information, the watermark grouping information will be embedded before the coded watermark. The grouping will be the training pattern for NN. Figure 8 illustrates the embedding process. First, the original image is divided into 8×8 blocks. Then each block will be DCT transformed from spatial to frequency domain. Next, if bit to embed is 0 then the anticipated output *AC*12" will be modified to –20; otherwise to 20. Finally, the image with embedded data will be IDCT back to spatial domain.

Figure 3. Encryption process for the watermark



Color image with embedded data

Figure 4. Extracting and enhancing the watermark with neural network

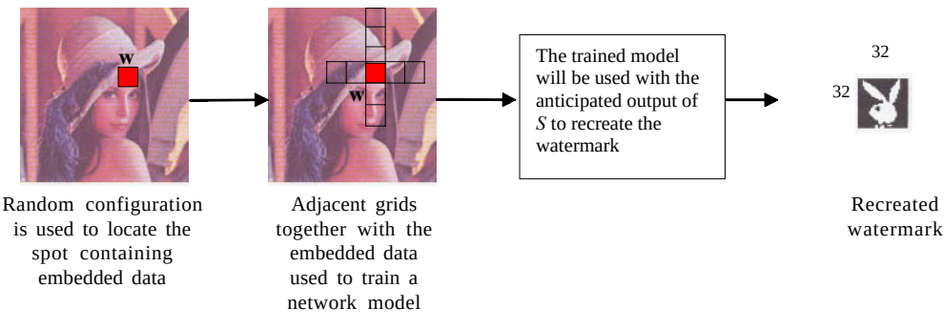
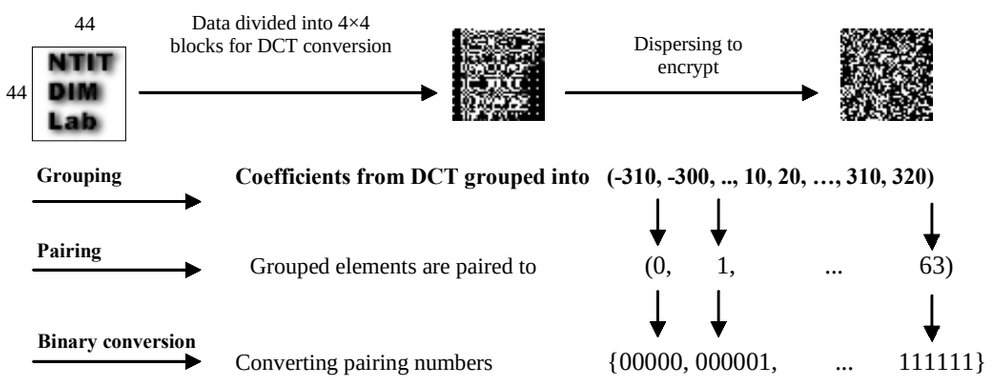

W
W
The trained model will be used with the anticipated output of S to recreate the watermark
32
32
Random configuration is used to locate the spot containing embedded data
Adjacent grids together with the embedded data used to train a network model
Recreated watermark

Figure 5. Original 1024×1024 grayscale image

Figure 6. The 44×44 grayscale watermark

Figure 7. Diagram on encrypting the watermark



The watermark will be extracted by using BPN to enhance the quality of the restored watermark. The training process to get the best output. As shown in Figure 9, the DC blocks containing the extra information were designated the {*AC*1", *AC*2", …, *AC*9"} training pattern. The trained output will help in identifying the 0's or 1's that were embedded. Each group of training patterns can be paired with an anticipated output. The sigmoid function (Hwang et al., 2000) for IDCT will require DC and AC values to fall between zero and one as shown in Eq. (1).

$$(c_i + 1000)/2000 \qquad (1)$$

where c is {DC, AC1", AC2", AC3", AC4", AC5", AC6", AC7", AC8", AC9", AC12"}, and j is the j^{th} neural node.

Figure 8. Diagram for watermark encryption

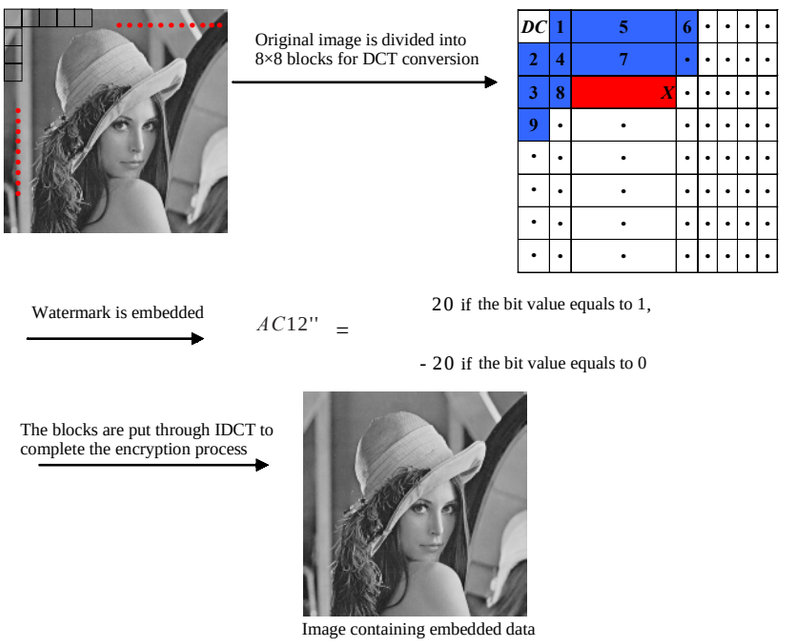
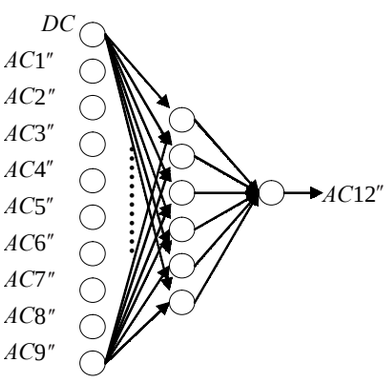


Figure 9. Diagram of the neural network for the proposed technique



Once NN training is completed, the blocks that actually contain the watermark {*DC*, *AC*1", *AC*2", …, *AC*9"},will be input variables to NN. The average value of the output computed through NN is defined by Eq.(2) and taken as the threshold value. As in Eq.(3), the embedded data will be 0 when the threshold is equal or greater than the computed *AC*12""; otherwise it is 1. Once the data in each block is extracted, a six-bit grouping is used to convert the extracted data into decimal and reverted back to between -310 and 320. The same random sequencing rule is used to recreate the dispersed watermark, which will be IDCT from the frequency to spatial domain. This completes the extraction process (see Figure 10).

$$\text{Threshold}=\sum \text{output value}/\text{amount of values}. \quad (2)$$

$$\text{extracted values of bit} = \begin{cases} 0 \text{ of Threshold} >= (c_i + 1000)/2000, \\ 1 \text{ if Threshold} < (c_i + 1000)/2000. \end{cases} \quad (3)$$

EXPERIMENTAL RESULTS AND ANALYSIS

A 1024×1024 grayscale host image and a 44×44 grayscale watermark will be used in the experiments. The Peak signal-to-noise ratio (PSNR) will be used to estimate the quality of the extracted digital watermark. An image quality of PSNR 30dB will be considered as acceptable if the image is visually acceptable, too.

Figure 10. Diagram for the watermark extraction by the improved method

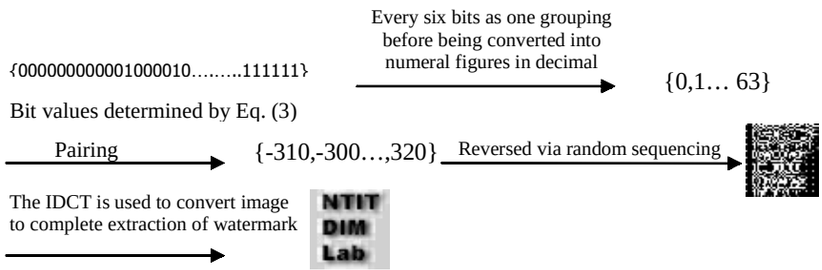

Every six bits as one grouping before being converted into numeral figures in decimal
{000000000001000010……..111111}
{0,1… 63}
Bit values determined by Eq. (3)
Pairing
{-310,-300…,320}
Reversed via random sequencing
The IDCT is used to convert image to complete extraction of watermark
NTIT
DIM
Lab

Figure 11. Comparing the original image with the embedded one

(a) Original image 1024×1024 (b) Embedded with data PSNR = 40.1 dB

Figure 12. Comparing the watermarks

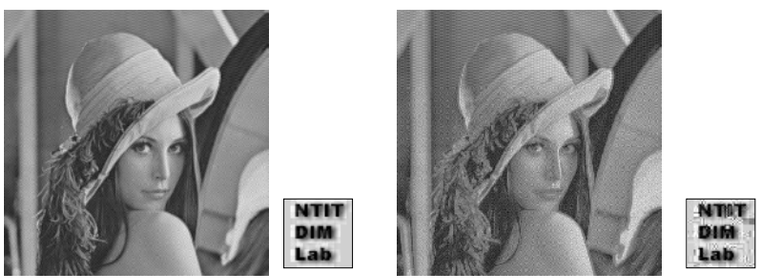
(a) Original 44×44 watermark

(b) Extracted watermark PSNR = 19.2 dB

Figure 13. Comparison of different levels of blur attack

(a) Blur attack oncePSNR 37.8dB PSNR 18.2dB (b) Blur attack twicePSNR 37.4 dB PSNR 16.4dB

Figure 14. Comparing different levels of sharpen attacks



(a)Sharpen twice PSNR 18.9 dB PSNR 19.2 dB (b)Sharpen thrice PSNR 11.3 dB PSNR 16.2 dB

When the image with embedded data in Figure 11(b) is compared with the original in Figure 11(a), very little difference is detected. Also, Figure 11(b) has a high PSNR of 40.1 dB. The watermark in Figure 12(b) is extracted from Figure 11(b). Although it has a low PSNR of 19.2 dB and showed some artifacts, it is still considered visually acceptable.

Severe Blur Attacks

The image with embedded data in Figure 13(a) is attacked once with blur while Figure 13(b) is attacked with blur twice. Their respective extracted watermarks show some artifacts, but are still considered as visually acceptable.

Severe Sharpen Attacks

Figure 14(a) illustrates the sharpen-twice attack, while Figure 14(b) is sharpened thrice. Their respective watermarks are successfully extracted but showed some artifacts. However, the wordings on the watermarks are clear. Therefore the extracted watermarks are considered as visually acceptable.

Lossy Compression Attacks

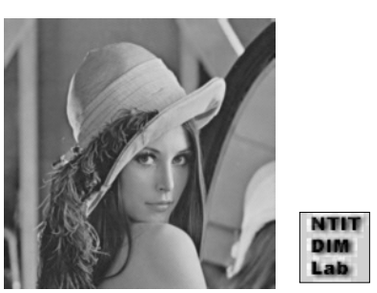
The images in Figures 15(a), (b), (c) and (d) are attacked with different rates of JPEG lossy compression. Their respective extracted watermarks showed the most distortions in the higher compression rate for Figure 15 (d). However, the wordings on the watermark are still recognizable.

CONCLUSION

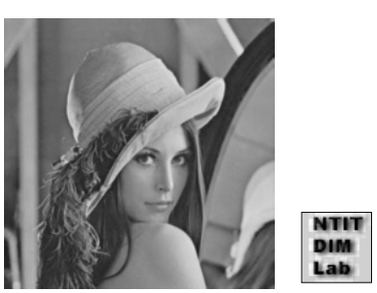
From the above experiments, the host image was attacked with different levels of blur, sharpen and JPEG lossy compression. In each case, the extracted watermark has some distortions. However, in all cases the most significant information (wordings) is intact and visually acceptable. Therefore, the watermark is robust. Furthermore, all the images have PSNRs well-above 30dBs and all visually similar to the original image. These proved that the proposed watermarking technique is robust and imperceptible.

As results have shown, the proposed improved method can effectively withstand attacks from com-

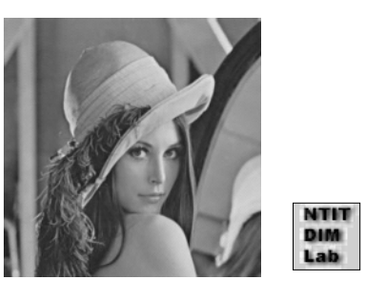
Figure 15. Comparing with different levels of lossy compression attack



(a) File size 164KB PSNR=38.9 dBCompression ratio 6.24:1 PSNR=19.2dB



(b) File size 112KBPSNR=39.3 dBCompression ratio 9.14:1PSNR=19.2dB



(c) File size 67.5KBPSNR=34.9 dBCompression ratio 15:1PSNR=19.2dB

(d)File size 50.2KB PSNR=38.5 dBCompression ratio 20:1PSNR=15.8dB

mon imaging processing. The watermark could be safely extracted to be used proof for rightful ownership. Its applications could be to protect copyrights of artworks on display on the internet, protection against illegal distributions and more. Future work could be to apply NN to retrain the extracted watermark to reduce distortions—the proposed method only used it to improve extraction. Another possible work could be to embed trace data as watermarks to prevent legal users from illegally redistributing data (Memon & Wong, 2001; Mukherjee, Maitra, & Acton, 2004).

REFERENCES

Cox, I.J. & Miller, M.L. (2001). Electronic watermarking: The first 50 years. *2001 IEEE Fourth Workshop on Multimedia Signal Processing,* October 3-5, (pp. 225-230).

Cox, I.J., Miller, M.L., & Bloom, J.A. (2000). Watermarking applications and their properties. *Proceedings on Information Technology: Coding and Computing*, March 27-29, (pp. 6-10).

Davis, K.J. & Najarian, K. (2001). Maximizing strength of digital watermarks using neural network. *Proceedings on Neural Network (IJCNN '01)*, July 15-19 (Vol. 4, pp. 2890-2898).

Du, J., Lee, C.H., Lee, H.K., & Suh, Y.H. (2002). BSS: A new approach for watermark attack. *Proceedings of the Fourth International Symposium on Multimedia Software Engineering*, December 11-13 (pp. 182-187).

Gonzalez, R.C. & Woods, R.E. (2002). *Digital image processing.* NJ: Prentice-Hall.

Hwang, M.S., Chang, C.C., & Hwang, K.F. (2000). Digital watermarking of images using neural networks. *Journal of Electronic Imaging*, *9*(4), 548-555.

Lin, C.Y., Wu, M., Bloom, J.A., Cox, I.J., Miller, M.L. & Lui, Y.M. (2001). Rotation, scale, and translation resilient watermarking for images. *IEEE Transactions on Image Processing*, *10*(5), 767-782.

Martin, H. & Kutter, M. (2001). Information retrieval in digital watermarking. *IEEE Communications Magazine*, *39*(8), 110-116.

Memon, N. & Wong, P.W. (2001). A buyer-seller watermarking protocol. *IEEE Transactions on Image Processing*, *10*(4), 643-649.

Miller, M.L., Doerr, G.J., & Cox, I.J. (2004). Applying informed coding and embedding to design a robust high-capacity watermark. *IEEE Transactions on Image Processing*, *13*(6), 792-807.

Mukherjee, D.P., Maitra, S., & Acton, S.T. (2004). Spatial domain digital watermarking of multimedia objects for buyer authentication. *IEEE Transactions on Multimedia*, *6*(1), 1-15.

Niu, X.M., Lu, Z.M., & Sun, S.H. (2000). Digital watermarking of still images with gray-level digital watermarks. *IEEE Transactions on Consumer Electronics*, *46*(1), 137-145.

Silva, R.A. & Mayer, J. (2003). Informed embedding for multibit watermarking. *XVI Brazilian Symposium on Computer Graphics and Image Processing* (SIBGRAPI 2003), October 12-15, (pp. 214-221).

Suhail, M.A. & Obaidat, M.S. (2001). On the digital watermarking in JPEG 2000. *The 8th IEEE International Conference on Electronics, Circuits and Systems, 2001 (ICECS)*, September 2-5 (Vol. 2, pp. 871-874).

Tsai, H.H., Cheng, J.S., & Yu, P.T. (2000). Digital watermarking based on Neural Networks for color images. *Elsevier on Signal Processing*, *81*, 663-671.

Zhang, J., Wang, N.C., & Xiong, F. (2002). A novel watermarking for images using Neural Networks. *Proceedings on Machine Learning and Cybernetics*, November 4-5, (Vol. 3, pp. 1405-1408).

KEY TERMS

Digital Watermark: An image or a logo in digital format embedded in a host image. The embedded data can later be used to prove the rightful ownership.

Grayscale Image: Shades of gray or continuous-tone of gray representing an image.

Image Copyrights: The rightful ownership of an image.

Imperceptible: Not easily detectable with the human visual system.

Neural network: A method that applies learning and relearning through a series of trials and errors to get the best performance.

Tamper: Making alterations to an image with unfriendly intent.

Robust: An image that can be successfully restored after suffering attacks from normal image processing or attacks of intent. The restored image must be visually acceptable.

Watermark Attack: Manipulation of the watermark by normal image processing or by making changes to with bad intents.

Digital Watermarking for Multimedia Security Management

Chang-Tsun Li
University of Warwick, UK

INTRODUCTION

The availability of versatile multimedia processing software and the far-reaching coverage of the interconnected networks have facilitated flawless copying and manipulations of digital media. The ever-advancing storage and retrieval technologies also have smoothed the way for large-scale multimedia database applications. However, abuses of these facilities and technologies pose pressing threats to multimedia security management in general, and multimedia copyright protection and content integrity verification in particular. Although cryptography has a long history of application to information and multimedia security, the undesirable characteristic of providing no protection to the media once decrypted has limited the feasibility of its widespread use. For example, an adversary can obtain the decryption key by purchasing a legal copy of the media but then redistributing the decrypted copies of the original.

In response to these challenges, digital watermarking schemes have been proposed in the last decade. The idea of digital watermarking is to embed a small amount of imperceptible secret information in the multimedia so that it can be extracted later for the purposes of copyright assertion, copy control, broadcasting, authentication, content integrity verification, and so forth. For example, a stream of binary bits generated, which identifies the owner of an image, can be taken as a watermark embedded at the least significant bit of the pixels or transformed coefficients by adjusting their value according to a predefined algorithm. Since the secret information is embedded in the content of the media, for the applications related to copyright protection where the watermark is intended to be robust, it does not get erased when the content is manipulated or undergoes format conversions. In this article, we will be addressing the main applications, security issues/challenges, solutions, and trends in the development of digital watermarking schemes. Bearing in mind that providing a comprehensive coverage of the applications, issues, and approaches of digital watermarking is not realistic due to the length limitation, we will refer the reader to some most recent publications in due course.

BACKGROUND

Unlike traditional watermarks on paper, which are visible to the eyes, digital watermarks can be designed to be imperceptible and removable. Throughout the rest of this article, the term *watermark*(*ing*) is used to refer to digital watermark(ing).

Various types of watermarking schemes have been proposed for different applications. For copyright-related applications, the embedded watermark is expected to be immune to various kinds of malicious and non-malicious manipulations to some extent, provided that the manipulated content is still valuable in terms of commercial significance or acceptable in terms of perceptual quality. Therefore, watermarking schemes for copyright-related applications are typically robust (Barni et al., 2002; Moulin & Ivanovic, 2003; Sebe & Domingo-Ferrer, 2003; Trappe et al., 2003); that is, they are designed to ignore or remain insensitive to manipulations.

Conversely, in medical, forensic, and intelligence or military applications, where content integrity and source authentication are a major concern, more emphases are placed on the schemes' capability of detecting forgeries and impersonations. Therefore, schemes of this type are usually fragile or semi-fragile and are intended to be intolerant to manipulations (Barreto et al., 2002; Li, 2004a; Li & Yang, 2003; Wong & Memom, 2000; Xie & Arce, 2001). Although a watermark is designed to be imperceptible to humans, the embedding is certainly intrusive and incurs distortion to the content.

In some authentication applications where any tiny changes to the content are not acceptable, the embedding distortion has to be compensated for perfectly.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

In an attempt to remove the watermark so as to completely recover the original media after passing the authentication process, reversible watermarking schemes have been proposed in the last few years (Alattar, 2004; Fridrich et al., 2002; Li, 2004b; Tan, 2003).

Requirements of digital watermarking vary across applications. The main requirements are low distortion, high capacity, and high security. One issue is that meeting all the three requirements simultaneously is usually infeasible; thus, trade-offs are frequently made to optimize the balance for each specific application. In many applications, where original media are not available at the watermark decoder, blind detection of the watermark without any prior knowledge about the original is desirable.

WATERMARKING SCHEMES AND THEIR APPLICATIONS

Digital watermarking schemes can be broadly classified into four categories: robust, fragile, semi-fragile, and reversible. While imperceptibility, low embedding distortion and security are the common requirements of all classes, each different category of scheme has different characteristics and, thus, is suitable for different applications. For example, while robustness is an essential requirement for copyright applications, it has no role in most authentication applications.

Robust Watermarking Schemes

Watermarks of robust schemes are required to survive manipulations, unless they have rendered the content valueless in some sense. This class of schemes has found its applications in the following areas. (The reader is reminded that the following list is not intended to be exhaustive, but just to identify some possible applications of multimedia security management.)

- **Ownership Proof and Identification:** A watermark containing the identification information of the content owner can be embedded in the host media for proving or identifying copyright ownership. However, proving ownership requires a higher level of security than ownership identification. For example, as pointed out by Craver et al. (1998), Bob could embed his watermark or make it appear that his watermark were embedded in a media owned and watermarked by Alice and could claim that this media belongs to him. In this scenario, the media contains both watermarks of Bob and Alice. Possible solutions to this problem of ambiguous ownerships have been reported in Craver et al. (1998) and Liu and Tan (2002).
- **Transaction Tracking/Fingerprinting:** The copyright owner could insert a unique watermark, which, for example, identifies the recipient, into each copy of the media and use it to trace the source, should illegal redistribution occur. The main challenge fingerprinting schemes face is the so-called *collusion attack* in which several legal copies of the same media are obtained to produce an approximation of the original unwatermarked version for illegal redistribution. Some recent proposals for tackling collusion attack can be found in Trappe et al. (2003) and Sebe et al. (2003).
- **Copy Control/Copy Prevention:** Illegal copying or recording is another common piracy scenario. One possible solution is to embed a never-copy watermark, which, when detected by the detector installed in the recording device, disallows further recording. However, this mechanism requires every recording device to have a watermark detector. It is difficult to persuade consumers to pay more for a device that restricts their freedom to make copies. This commercially undesirable requirement is unlikely to be met without the support of global legislation. The reader is referred to Bloom et al. (1999) for more details.
- **Broadcast Monitoring:** In advertisement applications, by embedding a watermark that is to be broadcast along with the host media, the advertisers can monitor whether or not the commercials they have paid for are aired by the broadcasters according to the contracts. More details can be found in De Strycker et al. (2000).

There are two major approaches to the designing of robust watermarking schemes; namely, spread spectrum (SS) watermarking (Cox et al., 1997) and

quantization index modulation (QIM) watermarking (Chen & Wornell, 2001). The idea behind SS-based schemes (Barni et al., 2002; Moulin et al., 2003) is to treat the watermark as a narrow-band signal and embed each bit in multiple samples of the host media, which is treated as a wide-band signal. The common approach taken by QIM-based schemes (Chen et al., 2001; Eggers et al., 2003; Liu & Smith, 2004) is first to establish an association between a set of watermarks and another set of quantizers with their codebooks predefined according to the watermarks. Then to embed a watermark, a set of features is extracted from the host media and quantized to the nearest code of the quantizer corresponding to the watermark. For both types of schemes, a common practice for ensuring low distortion and reducing the interference between the watermark and the host media is the so-called informed embedding, in which the information about the host media is exploited by the embedder (Cox et al., 2002).

Fragile Watermarking Schemes

In contrast to robust watermarking, fragile watermarks are sensitive to all kind of malicious and non-malicious manipulations (i.e., when manipulated, the watermarks are expected to be completely destroyed). Therefore, they are useful for the following applications:

- **Authentication:** In the areas of military intelligence and news broadcasting, authenticity of media sources is a key concern. By embedding a fragile watermark that identifies the source or producer in the media, the legitimate recipients of the marked media would be able to verify the authenticity of the received media by checking the presence of the source's or the producer's watermark. If the marked media is manipulated, the embedded watermark will become undetectable, and the recipient thus will know that the media is not trustworthy.
- **Content-Integrity Verification:** In the areas of medical image archiving, media recording of criminal events, and accident scene capturing for insurance and forensic purposes, content integrity may have a decisive impact on court rulings. The very presence of a fragile watermark in the original media allows the relevant parties to verify the integrity of the content.

An effective fragile watermarking scheme must have the capability of thwarting the attacks, such as cut-and-paste (i.e., cutting one region of the media and pasting it somewhere else in the same or another media) and vector quantization (i.e., forging a new marked image by combining some regions taken from different authenticated media while preserving their relative positions (Holliman & Memon, 2000). Some recent fragile schemes can be found in Li et al. (2003), Barreto et al. (2002), and Wong et al. (2000).

However, fragile watermarks are sensitive not only to malicious manipulations, but also to content-preserving operations such as lossy compression, transcoding, bit rate scaling, and frame rate conversion. Unfortunately, those content-preserving operations are sometimes necessary in many Internet and multimedia applications, making fragile watermarking feasible only in applications such as satellite imagery, military intelligence, and medical image archiving.

Semi-Fragile Watermarking Schemes

To facilitate the authentication and content-integrity verification for multimedia applications where content-preserving operations are a common practice, semi-fragile watermarking schemes have been proposed in the last few years (Ho & Li, 2004; Lin & Chang, 2000; Xie et al., 2001). This class of watermark is intended to be fragile only when the manipulations on the watermarked media are deemed malicious by the schemes. Usually, to achieve semi-fragility, the schemes exploit properties of or relationships among transformed coefficients of the media. Such properties and relationships are invariant to content-preserving operations while variant to malicious manipulations. The watermark is embedded by quantizing or adjusting the coefficients according to the watermark. The defined quantization step governs the fragility or sensitivity to manipulations and the degree of distortion.

However, an immediate result of coefficient quantization is that a unique watermark may be extracted from many different media that might have been

subjected to some form of content-preserving operation or malicious manipulation. Such a one-to-many correspondence can be problematic in terms false positives (i.e., a watermark that was never embedded is detected by the detector) and false negatives (i.e., the detector fails to detect an embedded watermark). Unfortunately, no optimal criteria for maintaining low false-positive and false-negative rates are currently in existence. Another challenge semi-fragile schemes face is how to distinguish content-preserving operations from malicious attacks. For example, transcoding may be deemed acceptable for one application, while it may be seen as malicious for another. Therefore, with these two issues, semi-fragile watermarking usually is not suitable for applications concerning legal and national security issues.

Reversible Watermarking Schemes

One limitation of the previously mentioned authentication schemes is that the distortion inflicted on the host media by the embedding process is permanent. Although the distortion is often insignificant, it may not be acceptable for some applications. For example, any tiny distortion of an image, even if it were a result of the watermark embedding process itself, in the legal cases of medical malpractice would cause serious debate on the integrity of the image. Therefore, it is desirable that watermarking schemes are capable of perfectly recovering the original media after passing the authentication process. Schemes with this capability often are referred to as *reversible watermarking schemes* (Alattar, 2004; Li, 2004b; Tian, 2003), also known as *invertible* (Fridrich et al., 2002) or *erasable watermarking* (Cox et al., 2002).

Taking a gray-scale image as an example, a common practice taken in Alattar (2004), Tian (2003), and Fridrich et al. (2002) is to look for two unequally represented sets of pixel groups such that changing the intensity of the elements belonging to one set changes its membership, making it belong to another set. A binary location map is then created, with each bit corresponding to one pixel group and the value (either 0 or 1) representing the membership of that pixel group. The location map subsequently undergoes some form of lossless compression so that its compressed version can be combined with the watermark, the actual payload, to form a bit stream for embedding. The embedding is carried out by changing the intensity of the pixel groups in order to make their membership consistent with the binary value of their corresponding bit in the bit stream. The extraction is simply a process of checking the membership of each pixel group of the watermarked image. If the image passes the authentication process, the original image can be recovered by uncompressing the location map and then changing the intensity of each pixel group so that its intensity becomes compatible with its actual membership recorded in the location map.

One limitation of all three schemes is that the ratio of the number of members in the two sets is highly dependent on the host image. Usually, images with more details or high-frequency components tend to have lower ratio, making the location map less compressible, and thus lowering the embedding capacity of the payload. An interesting scheme with media-independent embedding capacity is reported in Li (2004b) to alleviate this drawback.

FUTURE TRENDS

Quantization index modulation (QIM) schemes usually have higher embedding capacity than spread spectrum schemes and, therefore, are likely to be the dominating theme of research. Reversibility with media-independent embedding capacity will also be in the research agenda in the future for authentication applications. Although some perceptual models (De Vleeschouwer et al., 2002; Kutter & Winkler, 2002) have been proposed to ensure low embedding distortion, how distortion and robustness could be optimized is still an open question, and we expect new models will be proposed in the future.

Apart from security-oriented applications, which will continue to attract research interests, digital watermarking has been proved to be useful for broadcast monitoring, and we believe that it can be useful for other non-security-oriented applications such as error concealment and metadata hiding within multimedia content for legacy systems, so that the metadata can survive format conversions. The latter is particularly useful for document identification, as it allows us to reassociate medical images with patients' records and link multimedia to the World Wide Web.

CONCLUSION

Digital watermarking provides more options and promises for multimedia security management. However, despite its potentials, it is by no means a cure-all solution for multimedia security management. The solutions are more likely to remain application-dependent, and trade-offs between the conflicting requirements of low distortion, high capacity, complexity, and robustness still have to be made. Before trustworthiness can be evaluated, possible attacks for specific applications have to be studied at the development stage. For some applications such as copy control, non-technical backing by legislation is also crucial. With so many challenges and potential, we expect that digital watermarking will continue to be an active research area.

REFERENCES

Alattar, A.M. (2004). Reversible watermark using the difference expansion of a generalized integer transform. *IEEE Transactions on Image Processing, 13*(8), 1147-1156.

Barni, M., Bartolini, F., & Piva, A. (2002). Multichannel watermarking of color images. *IEEE Transactions on Circuits and Systems for Video Technology, 12*(3), 142-156.

Barreto, P.S.L.M., Kim, H.Y., & Rijmen, V. (2002). Toward secure public-key blockwise fragile authentication watermarking. *IEE Proceedings—Vision, Image and Signal Processing, 148*(2), 57-62.

Bloom, J.A., et al. (1999). Copy protection for DVD video. *Proceedings of the IEEE, 87*(7), 1267-1276.

Chen, B., & Wornell, G. W. (2001). Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, *47*(4), 1423-1443.

Cox, I., Miller, M., & Jeffrey, B. (2002). *Digital watermarking: principles and practice*. San Francisco, CA: Morgan Kaufmann.

Cox, I.J., Kilian, J., Leighton, F.T., & Shamoon, T. (1997). Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing, 6*(12), 1673-1687.

Craver, S., Memon, N., Yeo, B.-L., & Yeung, M.M. (1998). Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications. *IEEE Transactions on Selected Areas in Communications*, *16*(4), 573-386.

De Strycker, L., et al. (2000). Implementation of a real-time digital watermarking process for broadcast monitoring on a TriMedia VLIW processor. *IEE Proceedings—Vision, Image and Signal Processing, 147*(4), 371-376.

De Vleeschouwer, C., Delaigle, J.-F., & Macq, B. (2002). Invisibility and application functionalities in perceptual watermarking: An overview. *Proceedings of the IEEE , 90*(1), 64-77.

Eggers, J.J., Bauml, R., Tzschoppe, R., & Girod, B. (2003). Scalar Costa scheme for information embedding. *IEEE Transactions on Signal Processing, 51*(4), 1003-1019.

Fridrich, J., Goljan, M., & Du, R. (2002). Lossless data embedding—New paradigm in digital watermarking. *EURASIP Journal of Applied Signal Processing, 2002*(2), 185-196.

Ho, C.K., & Li, C.-T. (2004). Semi-fragile watermarking scheme for authentication of JPEG images. *Proceeding of the IEEE international Conference on Information Technology: Coding and Computing, I,* Las Vegas, Nevada.

Holliman, M., & Memon, N. (2000). Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes. *IEEE Transactions on Image Processing, 9*(3), 432-441.

Kutter, M., & Winkler, S. (2002). A vision-based masking model for spread-spectrum image watermarking. *IEEE Transactions on Image Processing, 11*(1), 16-25.

Li, C.-T. (2004a). Digital fragile watermarking scheme for authentication of JPEG images. *IEE Proceedings—Vision, Image, and Signal Processing, 151*(6), 460-466.

Li, C.-T. (2004b). *Reversible watermarking scheme with image-independent embedding capacity re-*

search report CS-RR-401. Coventry, UK: University of Warwick.

Li, C.-T., & Yang, F.-M. (2003). One-dimensional neighbourhood forming strategy for fragile watermarking. *Journal of Electronic Imaging, 12*(2), 284-291.

Lin, C.-Y., & Chang, S.-F. (2000). Semi-fragile watermarking for authenticating JPEG visual content. *Proceeding of the SPIE Conference on Security and Watermarking of Multimedia Contents II*, San Jose, CA.

Liu, R., & Tan, T. (2002). An SVD-based watermarking scheme for protecting rightful ownership. *IEEE Transactions on Multimedia, 4*(1), 121-128.

Liu, Y., & Smith, J.O. (2004). Watermarking sinusoidal audio representations by quantization index modulation in multiple frequencies. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Quebec, Canada.

Moulin, P., & Ivanovic, A. (2003). The zero-rate spread-spectrum watermarking game. *IEEE transactions on Signal Processing*, *51*(4), 1098-1117.

Sebe, F., & Domingo-Ferrer, J. (2003). Collusion-secure and cost-effective detection of unlawful multimedia redistribution. *IEEE Transactions on Systems, Man and Cybernetics, Part C, 33*(3), 382-389.

Tian, J. (2003). Reversible data embedding using a difference expansion. *IEEE Transactions on circuits and Systems for Video Technology*, *13*(8), 890-896.

Trappe, W., Wu, M., Wang, Z.J. & Liu, K.J.R. (2003). Anti-collusion fingerprinting for multimedia. *IEEE Transactions on Signal Processing, 51*(4), 1069-1087.

Wong, P.-W., & Memom, N. (2000). Secret and public key authentication watermarking schemes that resist vector quantization attack. *Proceeding of the SPIE Conference on Security and watermarking of Multimedia Contents II*, San Jose, CA.

Xie, L., & Arce, G.R. (2001). A class of authentication digital watermarks for secure multimedia communication. *IEEE Transactions on Image Processing*, *10*(11), 1754-1764.

KEY TERMS

Fragile Watermarking: A method for embedding a secret message that is intended to be undetectable even after minor malicious or non-malicious manipulations on the host Media in which it is embedded.

Informed Embedding: A data embedding method that exploits the information about the host Media available at the embedder side.

Lossy Compression: A multimedia compression operation that reduces the size of the media by removing data redundancy or discarding some details. The distortion incurred by the Operation is permanent.

Quantization Index Modulation (QIM) Watermarking: Given a set of watermarks and a set of quantizers with their codebooks predefined according to the watermarks, Quantization Index Modulation watermarking is a method of embedding a watermark in which a set of features are extracted from the host media and quantized to the nearest code of the quantizer corresponding to the watermark. The security resides in the secrecy of the Codebooks.

Reversible Watermarking: A watermarking method that allows the original host media to be Perfectly recovered after the marked media passes the authentication process.

Robust Watermarking: A method for embedding a secret message/watermark that is intended to be detectable even after significant malicious or non-malicious manipulations on the host Media in which it is embedded.

Semi-Fragile Watermarking: A method for embedding a secret message that is intended to be undetectable only after malicious manipulations of the host media in which it is embedded and detectable after non-malicious manipulations.

Spread Spectrum (SS) Watermarking: A method of embedding a watermark, a narrow-band signal, by spreading each bit of the watermark over several samples of the host media, a wide-band signal. The security resides in the secrecy of the spreading function.

Distance Education Delivery

Carol Wright
Pennsylvania State University, USA

INTRODUCTION

The term distance education is used to describe educational initiatives designed to compensate for and diminish distance in geography or distance in time. The introduction of technology to distance education has fundamentally changed the delivery, scope, expectations, and potential of distance education practices. Distance education programs are offered at all levels, including primary, secondary, higher, and professional education. The earliest antecedents of distance education at all levels are found worldwide in programs described most commonly as correspondence study, a print-dependent approach prolific in geographic areas where distance was a formidable obstacle to education. As each new technology over the last century became more commonly available, it was adopted by educational practitioners eager to improve communication and remove barriers between students and teachers.

BACKGROUND

Each developmental stage of technology incorporated elements of the old technology while pursuing new ones. Thus, early use of technology involved telephone, television, radio, audiotape, videotape, and primitive applications of computer-assisted learning to supplement print materials. The next iteration of distance education technologies, facilitating interactive conferencing capabilities, included teleconferencing, audioteleconferencing, and audiographic communication. Rapid adoption of the Internet and electronic communication has supported enhanced interactivity for both independent and collaborative work, access to dynamic databases, and the ability for students to create as well as assimilate knowledge. The rapid and pervasive incorporation of technology into all levels of education has been to a significant degree led by those involved in distance education. Virtual universities have evolved worldwide to offer comprehensive degrees. Yet, the technological advances are a threat to those who find themselves on the wrong side of the digital divide.

As distance delivery programs have increasingly incorporated technology, the term distance education has been used to distinguish them from more traditional, non-technology-based correspondence programs. As traditional resident higher education programs have adopted many of the technologies first introduced in distance education programs, the strong divisions between distance and resident programs have become increasingly blurred and have resulted in growing respect for distance education programs. In postsecondary education, technology-based distance education has gradually evolved into a profitable and attractive venture for corporations, creating strong competition for academic institutions. The involvement of the for-profit sector in the delivery of technical, professional, and academic degrees and certificates has, in turn, been a driving force in the renewed discussion of perennial higher education academic issues such as the nature of the learning and teaching experience; educational assessment; academic and professional accreditation; the delivery of student support services such as libraries, computing, and counseling services; and faculty issues such as promotion and tenure, workload, and compensation.

DISTANCE EDUCATION APPLICATIONS

In the primary and secondary environment, distance education is a successful solution for resource sharing for school districts unable to support specialized subject areas, students with mental or physical disabilities who are temporarily or permanently homebound, students with difficulties in a traditional classroom environment, repeat students in summer-school classes, advanced-placement students who

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

are able to access college-level programs, adults seeking to complete GED requirements, and the increasing numbers of families who choose a home-schooling option.

In the college and university environment, distance education is an attractive option for adult and nontraditional students, students who need to be away from campus for a semester, or those who have difficulties scheduling required courses in resident programs. Distance education delivery options have become a common dimension of almost all traditional institutions. For-profit entities are becoming a dominant force in the distance education arena as education evolves into a commodity, especially for advanced professional education and training, because of their ability to target the marketplace. With the certain need for continuing education and training across government, industry, business, higher education, and health care; the increasing affordability of technologies; and the growing demand for "just-in-time," on-demand delivery, distance education promises to be the answer for those who want and need the learning experience and necessary content delivered to their desktops at home or at their place of employment.

TECHNOLOGIES SUPPORTING DISTANCE LEARNING

Distance technologies involve transmitting combinations of voice, video, and data. The amount of bandwidth available determines the transmission capacity. More expensive, large-bandwidth systems include microwave signals, fiber optics, or wireless systems. Advanced distance education technologies include network infrastructures, real-time protocols, broadband and wireless communication tools, multimedia-streaming technology, distributed systems, mobile systems, multimedia-synchronization tools, intelligent tutoring, individualized distance learning, automatic FAQ (frequently asked question) reply methods, and copyright-protection and authentication mechanisms.

The network architecture determines the extent and flexibility of delivery. Discrete systems for Web support, course postings, course delivery, collaboration, discussion, and student support services are being replaced by Web-based learning-management or course-management systems that fully integrate all dimensions of the teaching-learning experience. These systems are supported by a network of networks that include hardware, software applications, and licensing; they connect intranets and off-campus, regional, national, and international networks. Wireless networks are rapidly expanding on multiple levels, including smaller personal-area networks with increased speed, wireless local-area networks (WLANS/WiFi) that serve confined spaces such as office buildings or libraries, metropolitan-area networks (WMANs) that connect buildings over a broader geographic area, and third-generation wireless cellular voice infrastructure that can transmit data. Internet 2 is a consortium of 206 universities in partnership with industry and government to develop and deploy advanced network applications and technologies, and it is a primary factor in the implementation of technological advances in distance and higher education. Another initiative, National LambdaRail (NLR), is composed of U.S. research universities and private-sector technology companies to provide a national-scale infrastructure for research and experimentation in next-generation networking technologies and applications, and to solve challenges of network architecture, end-to-end performance, and scaling.

Distance education delivery systems are commonly divided into two broad types: synchronous or asynchronous. Synchronous delivery requires that all participants—students, teachers, and facilitators—be connected at the same time with the ability to interact, transmit messages, and respond simultaneously. Online chat, interactive audio, or videoconferencing provide real-time interaction. The requirement that all participants come together at the same time, however, increases time constraints and decreases individual flexibility. Asynchronous delivery defines the anytime, anywhere experience where all participants work independently at times convenient to them, and it includes methods such as online discussion boards, e-mail, and video programming. The absence of immediate interaction with the teacher or other students is often criticized because of the isolation of participants, but this is acceptable for certain content areas and for adult or self-motivated learners. Sophisticated course design often seeks to integrate elements of both synchronous and asynchronous methods to meet individual needs and

course goals. The selection of effective technologies must focus on instructional outcomes, the needs of the learners, the requirements of the content, and internal and external constraints. Typically, this systematic approach will result in a mix of media, each serving a specific purpose.

Multimedia tools are critical because of the need to compensate for the lack of face-to-face contact. Distance education uses media for dual purposes—to deliver information and convey subject content—and also to facilitate communication between students and teachers. This attention to emerging media has meant that distance education has often taken a leadership position in the adoption of technology and multimedia by the broader educational community. The definition of multimedia continually changes since new applications and technological advances result in a constantly evolving array of hardware and software, which allow audio and visual data to be combined in new ways. The personal computer, the Internet, authoring and editing software, and newer media such as wireless personal devices have created dynamic digital learning environments that facilitate interactivity, autonomous learning, assessment opportunities, and virtual learning communities. Multimedia packages that consist of suites of software applications facilitate the integration of state-of-the-art communication, collaboration, content-delivery, student-assessment, and course-management capabilities.

EVALUATION OF DISTANCE EDUCATION PROGRAMS

Comprehensive evaluation must be an integral component of distance education programs. SWOT analysis, a critical component of the strategic planning process, is an effective tool that helps to identify resources and capabilities, and to formulate strategies to accomplish goals. SWOT involves a scan of the internal and external environment, and identifies internal environmental factors as strengths (S) or weaknesses (W), and external factors as opportunities (O) or threats (T).

Early efforts to evaluate distance education focused on the transfer of course content and found that, compared to traditional course delivery and face-to-face instruction, there is no significant difference.

Future evaluation should examine more substantive and fundamental questions, such as the success in meeting stated learner outcomes, student-to-student interactions, teacher feedback, the development of learning communities, the incorporation of various learning styles, the development of effective teacher-training programs, the degree to which courses and programs are recognized in professional and employment arenas, the transferability of coursework across institutions, and enrollment and course-completion rates.

FUTURE TRENDS: ISSUES AND CHALLENGES

Among the continuing challenges for distance education are online ethics, intellectual property and copyrights, faculty issues, institutional accreditation, financial aid, and student support services.

Ethics in the Online Environment

Ethical behavior and academic honesty among students is of concern in any educational environment, and the online distance environment lends itself to significant abuse. Strategies to discourage and identify such behaviors require advance planning and aggressive attention. Course design, teaching techniques, and subscriptions to online services that help faculty detect plagiarism can be effective. Some useful approaches include designing assignments that are project based and focus on a task resulting in a product, and that require some degree of cooperation and coordination among students. Such products should incorporate students' own experiences and emphasize the process rather that simply the end result. Assignments can rotate across different semesters so that they are less predictable. Assignments that consist of small, sequential, individualized tasks can ensure that students keep up with class readings and respond to class assignments. High levels of instructor and student interaction, frequent e-mail contact, and online chats can ensure participation. An electronic archived record of all correspondence permits the tracking of content and variations in a student's writing style. All

courses should include an academic integrity policy. In an electronic environment where downloading and cut-and-paste are routine habits of information gathering, instructors must directly address ethical issues concerning the submission of such materials as a student's own work.

Intellectual Property and Copyright

Internationally, intellectual property and copyright issues are regulated primarily by the World Intellectual Property Organization (WIPO) and the European Union (EU). WIPO, including 180 member states, aims to ensure that the rights of creators and owners of intellectual property are protected worldwide. EU is concerned with these issues with the objectives of enhancing the functioning of the single market and harmonizing rules to insure uniform protection within the EU.

In a traditional classroom environment, faculty develop course materials, select appropriate readings, and develop a syllabus and curriculum for which they correctly claim intellectual property rights and ownership. Occasionally, this work is translated to textbooks for which faculty likewise maintain intellectual property rights. Conversely, in the online environment, institutions often claim either complete or partial ownership of the intellectual content because the work, when posted on the Internet, goes beyond the confines of the classroom; because online courses are often commissioned separately from standard employment contracts; and because the infrastructure supporting the transmission of the content is owned by the institution. The question of ownership is a divisive one, and debate continues; a resolution may found in varying formulas that divide royalties among faculty, departments and colleges, and research offices.

Prior to the TEACH Act of 2002 (Technology, Education and Copyright Harmonization Act), using copyright-protected materials in a self-contained classroom in the United States was within fair use, but posting the same materials on a Web page with potential worldwide distribution exceeded fair-use guidelines. The limitation posed a severe handicap on U.S. distance education programs. In November 2002, the TEACH Act generally extended to nonprofit, accredited institutions, for mediated instructional activities only, the same type of right to use copyright-protected materials that a teacher would be allowed to use in a physical classroom. TEACH expands existing exemptions to allow for the digital transmission of copyrighted materials, including through Web sites, so they may be viewed by enrolled students.

Faculty Issues

Whereas for-profit distance education institutions hire faculty with the express purpose of teaching specific courses, the climate and culture of traditional academic institutions often does not support distance education initiatives. Distance courses are frequently not included in a standard faculty workload, raising questions of faculty incentives and rewards. In cases where research institutions have promotion and tenure requirements that emphasize scholarship, service, and research at least as much as teaching, it is difficult for young faculty to commit to additional teaching assignments even if monetary compensation is provided. Even in universities and colleges where selected courses and programs are successful, limited institutional resources may prohibit program growth and diminish scalability.

Beyond the usual skills required of instructors, distance education faculty must meet additional expectations. They must develop an understanding of the characteristics and needs of distant students, become highly proficient in technology delivery, adapt their teaching styles to accommodate the needs and expectations of multiple and diverse audiences, and be a skilled facilitator as well as content provider. They therefore require strong institutional support for course design and delivery, technical support, and colleagues with whom to share common interests and concerns.

Financial Aid

Distance education students often have far fewer options for financial aid than do traditional students. Financial aid is often not available to students who are enrolled at school less than half time or who attend less than 30 weeks of instruction in an academic year; for courses that provide less than 12 weeks of instruction, examination, or preparation for examinations or that are not tied to standard course

lengths such as semesters or quarters; or for courses offered by institutions where more than 50% of the students are distance learners or more than 50% of the courses are offered by computer, correspondence, or video. Such regulations, developed to curb abuses exacerbated by Internet diploma mills, are in direct contradiction to the flexibility and advantages offered by distance programs. In 1988, the United States amended the 1965 Higher Education Act to support a distance education demonstration program, still in progress, intended to study the factors that define quality distance education experiences and to test the viability of increased financial support.

Accreditation

Accreditation has long been viewed as the vehicle to monitor the quality of educational institutions. Accreditation in countries outside of the United States is normally handled by ministries of education or other government entities. The Council for Higher Education Accreditation (CHEA), in conjunction with other higher education groups, is working to maintain and expand international accreditation and quality assurance. U.S. accreditation is offered through regional bodies or specialized professional or programmatic groups, and is complicated by overlaps between federal, regional, and state accrediting agencies. The rise of distance programs has increased the number of nationally accredited institutions, generally for-profit colleges and universities, whose students find that their courses are routinely not transferable to regionally accredited institutions. Students are sometimes able to persuade other schools to accept distance credits, but many do not. The dilemma demonstrates existing prejudice against distance education and is a serious deterrent to students, slowing the growth of online education. Recent discussion has suggested that the entire accreditation process be reviewed and restructured by the U.S. Department of Education.

Student Support Services

Distance students require many of the same academic support services offered to traditional students. Primary ones include academic advising and access to library and information resources. The professional associations for each of these areas (the Association of College and Research Libraries/ American Library Association and the National Academic Advising Association/Nacada) have developed standards to guide the delivery of quality service to distance students. Such guidelines assure equitable treatment and are a mechanism to measure quality for accreditation. The best designed courses and programs can fail without careful attention to executing the myriad details required for program success. Examples include application and admissions processes, student orientation, course registration processes, course drops or deferrals, placement examinations, computer technical support, financial-aid support, disability services, general student advocacy issues, materials duplication and distribution, textbook ordering, and securing of copyright clearances.

CONCLUSION

Distance education promises to become an increasingly pervasive and dominant force in educational delivery, accelerated by advancing communication and information technologies. It will help answer the demands for education within a digital information environment, the ever-increasing needs for continuing training on a global scale, and individual interest in lifelong learning. The expansion of distance education will likely force significant changes in the way more traditional education is delivered, and will in time be totally assimilated into the educational experience.

REFERENCES

Chien, C. (2003). Interactivity and interactive functions in Web-based learning systems: A technical framework for designers. *British Journal of Educational Technology, 34*(3), 265-279.

D'Antoni, S. (Ed.). (2004). *The virtual university models and messages: Lessons from case studies.* Paris: UNESCO/International Institute for Educational Planning. Retrieved August 13, 2004, from *http://www.unesco.org/iiep/virtualuniversity/home.php*

Discenza, R., Howard, C., & Schenk, K. D. (Eds.). (2002). *Design and management of effective distance learning programs.* Hershey, PA: Idea Group Publishing.

DiStefano, A., Rudestam, K., & Silverman, R. (Eds.). (2004). *Encyclopedia of distributed learning*. Thousand Oaks, CA: Sage Publications.

Heberling, M. (2002). Maintaining academic integrity in online education. *Online Journal of Distance Learning Administration, 5*(1). Retrieved August 13, 2004, from *http://www.westga.edu/%7Edistance/ojdla/spring51/herberling51.html*

Holzer, E. (2004). Professional development of teacher educators in asynchronous electronic environments: Challenges, opportunities and preliminary insights from practice. *Educational Media International, 41*(1).

Instructional Technology Council. (2004). *Distance education reports and abstracts*. Washington, DC: Author. Retrieved August 13, 2004 from *http://144.162.197.250/reports.htm#Costs%20for%20Distance%20Learning*

Internet 2. (2004). Retrieved August 13, 2004 from *http://www.internet2.edu/*

Johnston, J., & Toms Barker, L. (2002). *Assessing the impact of technology in teaching and learning: A sourcebook for evaluators*. Arbor, MI: University of Michigan, Institute for Social Research.

Lynch, M. M. (2002). *The online educator: A guide to creating the virtual classroom*. New York: Routledge Falmer.

Moore, M. G., & Anderson, W. G. (Eds.). (2003). *Handbook of distance education.* Mahwah, NJ: Lawrence Erlbaum Associates.

National LambdaRail. (2004). Retrieved August 13, 2004 from *http://www.nlr.net/*

Potashnik, M., & Capper, J. (1998, March). Distance education: Growth and diversity. *Finance & Development, 35*(1). Retrieved from *http://www.worldbank.org/fandd/english/0398/articles/0110398.htm*

Rovai, A. P. (2003). A practical framework for evaluating online distance education programs. *The Internet and Higher Education, 6*(2), 109-124.

Salmon, G. (2003). *E-moderating: The key to teaching and learning online* (2nd ed.). London: Routledge Falmer.

Taylor, J. C. (1999). *Distance education: The fifth generation*. Nineteenth ICDE World Conference on Open Learning and Distance Education, Vienna Austria. Retrieved August 13, 2004, from *http://www.usq.edu.au/users/taylorj/publications_ presentations/1999vienna_ 5thGeneration.doc*

Technology, Education and Copyright Harmonization Act. (2002). U.S. Copyright Office. Retrieved August 13, 2004 from *http://www.copyright.gov/legislation/pl107-273.html#13301*

Threlkeld, R., & Brzoska, K. (1994). Research in distance education. In B. Willis (Ed.), *Distance education: Strategies and tools*, (pp. 41-66). Englewood Cliffs, NJ: Educational Technology Publications, Inc.

Tiene, D. (2002). Addressing the global digital divide and its impact on educational opportunity. *Educational Media International, 39*(3-4), 211-222.

Tiffin, J., & Rajasingham, L. (2003). *The global university.* London: Routledge Falmer.

KEY TERMS

Asynchronous Distance Delivery: An anytime, anywhere experience where all participants work independently at times convenient to them and that include methods such as online discussion boards, e-mail, and video programming, and the implicit absence of immediate interaction with the teacher or other students.

Audiographic Communication: A multimedia approach with simultaneous resources for listening, viewing, and interacting with materials.

Audioteleconferencing: Voice-only communication via ordinary phone lines. Audio systems

include telephone conference calls as well as more sophisticated systems that connect multiple locations.

Synchronous Distance Delivery: Requires that all involved—students, teachers, and facilitators—be connected and participating at the same time with the ability to interact and to transmit messages and responses simultaneously.

Teleconferencing: Communication that allows participants to hear and see each other at multiple remote locations.

Virtual Universities: Institutions that exclusively offer distance courses and programs, often on a global scale.

Web Conferencing: Communication that allows audio participation with simultaneous visual presentation through a Web browser.

Distanced Leadership and Multimedia

Stacey L. Connaughton
Purdue University, USA

INTRODUCTION

At the dawn of the 21st century, more and more organizations in various industries have adopted geographically dispersed work groups and are utilizing advanced technologies to communicate with them (Benson-Armer & Hsieh, 1997; Hymowitz, 1999; Townsend, DeMarie & Hendrickson, 1998; Van Aken, Hop & Post, 1998). This geographical dispersion varies in form. For example, some organizations have adopted "telecommuting," in which members may work at home, on the road and/or at the office (Hymowitz, 1999). Other organizations have created teams that are globally dispersed. A leader located in Palo Alto, California, for example, may be responsible for coordinating employees in Belgium, China and Mexico.

This article examines the role of communication and multimedia in leading people across time and space. To do so, I first note the significance of distanced work relationships; then, outline various conceptualizations of "distance" evident in the literature; next, discuss the role of multimedia in those relationships; and conclude by forecasting future trends. Throughout the article, the term "distanced leadership" is used to refer to leadership in geographically dispersed contexts.

THE PROLIFERATION OF DISTANCED LEADERSHIP

New organizational forms have become increasingly prevalent in recent years. Indeed, many contemporary organizations and teams span time and space. Physical separation of organizational and/or team members is a defining characteristic of virtual organizations and teams (Jarvenpaa & Leidner, 1998; Majchrzak, Rice, King, Malhotra & Ba, 2000; Warkentin, Sayeed & Hightower, 1997; Wiesenfeld, Raghuram & Garud, 1999), geographically dispersed teams (Connaughton & Daly, 2003, 2004a, 2004b; Shockley-Zalabak, 2002), dispersed network organizations (Rosenfeld, Richman & May, 2004) and telework operations (Hylmo & Buzzanell, 2002; Leonardi, Jackson & Marsh, 2004; Scott & Timmerman, 1999). In these forms, the organization or team is constituted in its interaction and formal and informal networks. By 2005, 20% of the world's work force is expected to work virtually (Prashad, 2003). Indeed, scholars have called on leadership scholarship to "stretch its boundaries to match the elastic nature of global work" (Davis, 2003, p. 48).

Geographical dispersion affords organizations both opportunities and challenges to both business and communication. Table 1 summarizes these issues as they often appear in the literature.

On the one hand, geographically dispersed teams present organizations with many opportunities. They can help organizations maximize productivity and lower costs (Davenport & Pearlson, 1998). And, they can enable organizations to serve international customers and capitalize on globally dispersed talent (Majchrzak, Rice, King, Malhotra & Ba, 2000; Zaccaro & Bader, 2003). Ideally, this geographical dispersion is designed to foster productivity from, and cooperation among, organizational members, just as if they were co-located with one another (see Handy, 1995; Upton & McAfee, 1996).

Yet geographical dispersion also poses some challenges, specifically with regard to leadership. Previous research indicates that (a) a leader's "social presence" may be more difficult to achieve in distanced settings (Kiesler & Sproull, 1992; Warkentin, Sayeed & Hightower, 1997); (b) trust among leaders and team members may be swift yet fleeting (Jarvenpaa, Knoll & Leidner, 1998); (c) members' identification with the team, organization, and leader may be challenged over distance (Connaughton & Daly, 2004b); and (d) communication among leaders and team members may be complicated by diverse ethnic, communication and organizational backgrounds (Cascio, 1999; Cascio & Shurygailo, 2003).

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. Opportunities and challenges of globally dispersed teams

Opportunities	Challenges
• Reduce Costs	• Time Zone Differences
• Serve International Customers/Clients	• Varied Communication Norms
• Integrate Global Talent	• Language Differences
	• Limited Face-to-Face Contact

These challenges are put into perspective when one compares what may take place in physically proximate offices to what often happens in distanced work relationships. It has been suggested that co-located office settings provide more opportunities for organizational members to communicate frequently and spontaneously with each other; they allow for potential to interact immediately for troubleshooting; they foster a forum in which to directly access information; and they enable the development and maintenance of relationships (Davenport & Pearlson, 1998). Often, leaders who are co-located with their team members develop and energize relationships with their team through informal as well as formal interaction. In globally dispersed organizations, however, there may be fewer opportunities to informally communicate, leaving some distanced employees feeling isolated from their leaders and from events that take place at the central organization (Van Aken, Hop & Post, 1998; Wiesenfeld, Raghuram & Garud, 1998).

CONCEPTUALIZATIONS OF "DISTANCE"

Research on distanced work relationships, including that related to leadership, defines "distance" in different ways. Some scholars examine physical distance, when individuals and leaders are separated by geography (see Antonakis & Atwater, 2002; Kerr & Jermier, 1978). Other scholars investigate social or psychosocial distance, which often refers to perceived differences in status, rank, authority, social standing and power among leaders and followers, all of which may affect the intimacy and social interactions that take place between leaders and followers (see Antonakis & Atwater, 2002; Napier & Ferris, 1993).

Some researchers conceive of physical distance and social distance as related constructs, functioning in a similar manner (see Howell & Hall-Merenda, 1999). Others argue that physical distance and social distance are distinct and should be considered as separate constructs in research. Among them, Antonakis and Atwater (2002) also add a third dimension of distance, perceived interaction frequency, which they define as the perceived degree to which leaders interact with their followers. They propose that physical distance, social distance and perceived interaction frequency are measurable and are separate dimensions, each of which describes an element of "distance" in dispersed work relationships.

Other research examines how individuals perceive distance in geographically dispersed work contexts. For example, in a study of 46 teleworkers in a variety of industries, Leonardi, Jackson and Marsh (2004) argue that these individuals *manage* distance in various ways. The authors conclude that dispersed individuals do not all perceive distance similarly, and that they manipulate the fact that they are geographically distant from others in order to satisfy individual needs. In the authors' words, "… distance is much more than a mere outcome of the use of ICTs; it is rather a tool virtual team members can use to manage their relationships with their coworkers and their organizations" (p. 169).

MULTIMEDIA AND DISTANCED LEADERSHIP

The published work on multimedia, communication technologies and dispersed leadership can be grouped into two broad categories: that which discusses effective practices for using media to forge connections across time and space; and that which addresses key assumptions in previous research, particularly with respect to the perceived necessity of face-to-face interaction and to the impact physical distance has on work relationships.

Effective Practices

Some published work advances effective practices, highlighting various ways that leaders can utilize multiple media to foster connections with distanced

employees across time and space. Among the recommendations, researchers have noted: (a) the creation of Web sites, where project managers can post their "lessons learned" and share effective practices with leaders at other sites; (b) the utilization of electronic forums to advertise what "works" in the regions and to propagate those ideas to headquarters and other remote sites; and (c) the development of internal electronic bulletin boards (one devoted to leaders; another devoted to members), where project leaders and team members can ask questions and receive suggestions from other project leaders and members (see Burtha & Connaughton, 2004; Connaughton & Daly, 2003). Majchrzak, Malhotra, Stamps and Lipnack (2004) note that these virtual work spaces should be considered more than "networked drives with shared files" (p. 134). These virtual work spaces must be accessible to everyone at all times, and a place where the team is reminded of its mission, purpose, decisions and future objectives.

In addition to explaining effective practices, existing research advances propositions about which media function particularly well to achieve various leadership objectives. One of these works is based on a series of interviews with distanced leaders about what media they perceive to be effective in executing various leadership functions across time and space (Connaughton & Daly, 2003). Distanced leaders interviewed in this study perceive that face-to-face communication is optimal for achieving objectives, but acknowledge that it is not always possible when employees and team members are dispersed. The research findings suggest that face-to-face communication is best used to set vision, reach policy decisions and begin to build relationships. When face to face is not an option, regularly scheduled telephone calls are most effectively used to exchange important task-related information, maintain relationships, appraise performance and coordinate teams. And, electronic mail (e-mail) is most effective to exchange technical information, give specific directions, update interested parties and maintain relationships. (For further discussion of technologies and virtual contexts, see Contractor & Eisenberg, 1990; Ferris & Minielli, 2004; Haythornthwaite & Wellman, 1998; Majchrzak, Rice, King, Malhotra & Ba, 2000.)

Addressing Assumptions of Previous Research

Recent work on distanced leadership has begun to carefully consider two assumptions about working in dispersed contexts. Those assumptions are: (a) that face-to-face communication is related to organizational outcomes; and (b) that physical distance necessarily is an impediment to productive and satisfying work relationships.

- **Assumption:** *Face-to-face communication is critical*. It has been argued that, despite the existence of new media, face-to-face communication is still vitally important to achieving organizational outcomes (Cohen & Prusak, 2001). Some scholarship compares experiences of individuals working proximately with one another (and who can communicate face to face) with individuals working apart from one another. For instance, Warkentin, Sayeed and Hightower (1997) found that face-to-face group members perceive greater team cohesion, and more satisfaction with both the group interaction process and group outcomes than did their distanced counterparts. One conclusion that could be drawn from this research is that individuals prefer to work in close proximity to leaders.

  Zack (1994) and Alge, Wiethoff and Klein (2003) found, however, that although initial face-to-face interactions are quite helpful for teamwork, as time goes on and team members come to better understand one another, mediated communication such as e-mail could be used to accomplish tasks. Scholars are also beginning to explore the processes of teams who never meet face to face and yet still function (Bell & Kozlowski, 2002; Davis, 2003). Continued research in this area may challenge the assumption that face-to-face communication is a necessary ingredient of effective distanced work relationships.
- **Assumption:** *Physical distance necessarily challenges work relationships*. Previous research on distanced work relationships as-

sumes that physical distance complicates performance and leader-follower relationships because distance makes it difficult for leaders to engage in relational and task-related behaviors with followers (see Kerr & Jermier, 1978; Olson & Olson, 2000). Often, scholars contrast distanced leadership with proximate leadership, and claim that physical proximity enables more effective communication between leaders and followers (Yagil, 1998).

However, the perceived *accessibility* of people in the distanced relationship may matter in predicting important outcomes as well (Cascio & Shurygailo, 2003; Napier & Ferris, 1993). Perceived accessibility refers to the distanced employees' perception that they can contact or reach their leader when so desired. Indeed, previous research has suggested that frequent interaction is critical to establishing a feeling of connection across time and space (Antonakis & Atwater, 2002; Connaughton & Daly, 2004b; Leonardi et al., 2004). And, distance may be perceived in positively valenced ways. As Leonardi et al. (2004) have argued, distance may be strategically managed by some distanced leaders and employees to be an *opportunity* rather than a necessary impediment to work relationships.

FUTURE TRENDS

Thousands of companies in diverse industries now have distanced leaders (see Apgar, 1998; Bryan & Fraser, 1999; Hymowitz, 1999; Lipnack & Stamps, 1997; McCune, 1998). And these leaders face the complex task of managing people who are separated from organizational headquarters by time and space.

Future investigations should consider related organizational trends. For instance, does leadership of geographically dispersed ad hoc teams (that are assembled for short-term projects) differ from the type of distanced leadership described here? If so, how? How does one manage contractors and consultants (who may not have loyalty to the organization) from afar? And, given trends in international customer service, how do organizations effectively serve and lead customers from afar?

Future researchers should also continue to develop theoretical models of distanced leadership as well as continue to conduct empirical work on these and other variables. For instance, it will be important to investigate whether actual physical distance *per se* is the most essential defining feature of a dispersed relationship. Instead, perhaps *physical distance* and *access* to leaders and team members function together to affect relationships and outcomes.

Another important issue for both scholars and practitioners is the assumption made by many that distanced teams have more difficulty than face-to-face teams. That presumption warrants empirical testing. The leaders we have talked with in our research have been quite insistent that face-to-face exchanges offer them the optimal medium for communication. None of the leaders interviewed consider mediated technologies as being effective for handling personnel issues, conflicts and relational development. Yet a question arises: Are these responses tied to levels of experience and training with the technologies, generational differences or other factors? It may be that with more experience using various technologies for communication and more perceived expertise with them that people's preference for face-to-face communication for various tasks may diminish. Future research may find that some distanced employees actually prefer mediated communication with their leader.

CONCLUSION

As organizations become more global, as talent becomes more dispersed and as technologies enable people to do far more from afar, distanced leadership and dispersed work relationships will continue to be important to organizations in the 21st century. Given those trends, the issues discussed in this chapter will become ever more critical for scholars and practitioners to consider.

REFERENCES

Alge, B.J., Wiethoff, C., & Klein, H.J. (2003). When does the medium matter? Knowledge-building expe-

riences and opportunities in decision-making teams. *Organizational Behavior and Human Decision Processes, 91*, 26-37.

Antonakis, J., & Atwater, L. (2002). Leader distance: A review and a proposed theory. *Leadership Quarterly, 13*, 673-704.

Apgar, IV, M. (1998). The alternative workplace: Changing where and how people work. *Harvard Business Review, 76*(3), 121-136.

Bell, B.S., & Kozlowski, S.W.J. (2002). A typology of virtual teams: Implications for effective leadership. *Group & Organization Management, 27,* 14-49.

Benson-Armer, R., & Hsieh, T. (1997). Teamwork across time and space. *The McKinsey Quarterly, 4,* 18-27.

Bryan, L.L., & Fraser, J.N. (1999). Getting to global. *The McKinsey Quarterly, 4*, 28-37.

Burtha, M., & Connaughton, S.L. (2004). Learning the secrets of long-distance leadership: Eight principles to cultivate effective virtual teams. *Knowledge Management Review, 7*, 24-27.

Cascio, W.F. (1999). Virtual workplaces: Implications for organizational behavior. In C.L. Cooper & D.M. Rousseau (Eds.), *Trends in organizational behavior* (pp. 1-14). Chichester, UK: John Wiley & Sons.

Cascio, W.F. & Shurygailo, S. (2003). E-leadership and virtual teams. *Organizational Dynamics, 31*, 362-376.

Cohen, D., & Prusak, L. (2001). *In good company: How social capital makes organizations work*. Cambridge, MA: Harvard University Press.

Connaughton, S.L., & Daly, J.A. (2003). Long distance leadership: Communicative strategies for leading virtual teams. In D.J. Pauleen (Ed.), *Virtual teams: Projects, protocols, and processes* (pp. 116-144). Hershey, PA: Idea Group Publishing.

Connaughton, S.L., & Daly, J.A. (2004a). Leading from afar: Strategies for effectively leading virtual teams. In S. Godar & S.P. Ferris (Eds.), *Virtual and collaborative teams: Process, technologies & practice* (pp. 49-75). Hershey, PA: Idea Group Publishing.

Connaughton, S.L., & Daly, J.A. (2004b). Leading in geographically dispersed organizations: An empirical study of long distance leadership behaviors from the perspective of individuals being led from afar. *Corporate Communication: An International Journal, 9*(2), 89-103.

Contractor, N.S., & Eisenberg, E.M. (1990). Communication networks and new media in organizations. In J. Fulk & C. Steinfield (Eds.), *Organizations and communication technology* (pp. 143-172). Thousand Oaks, CA: Sage Publications.

Davenport, T.H., & Pearlson, K. (1998). Two cheers for the virtual office. *Sloan Management Review, 39,* 51-65.

Davis, D.D. (2003). The Tao of leadership in virtual teams. *Organizational Dynamics, 33*(1), 47-62.

Ferris, S.P. & Minielli, M.C. (2004). Technology and virtual teams. In S. Godar & S.P. Ferris (Eds.), *Virtual and collaborative teams: Process, technologies & practice* (pp. 193-211). Hershey, PA: Idea Group Publishing.

Handy, C. (1995, May-June). Trust and the virtual organization. *Harvard Business Review,* 40-50.

Haythornthwaite, C., & Wellman, B. (1998). Work, friendship, and media use for information exchange in a networked organization. *Journal of the American Society for Information Science, 49*, 1101-1114.

Howell, J.M., & Hall-Merenda (1999). The ties that bind: The impact of leader-member exchange, transformational and transactional leadership, and distance on predicting follower performance. *Journal of Applied Psychology, 84*, 680-694.

Hylmo, A., & Buzzanell, P.M. (2002). Telecommuting as viewed through cultural lenses: An empirical investigation of the discourses of utopia, identity, and mystery. *Communication Monographs, 69*, 329-356.

Hymowitz, C. (1999, April 6). Remote managers find ways to narrow the distance gap. *The Wall Street Journal*, B1.

Jarvenpaa, S., & Leidner, D.E. (1998). Communication and trust in global virtual teams. *Journal of*

Computer Mediated Communication, 3. Online from *www.ascusc.org/jcmc/vol3/issue4/jarvenpaa.html*

Jarvenpaa, S., Knoll, K., & Leidner, D.E. (1998). Is anybody out there? Antecedents of trust in global virtual teams. *Journal of Management Systems, 14,* 29-64.

Kerr, S., & Jermier, J.M. (1978). Substitutes for leadership: Their meaning and measurement. *Organizational Behavior and Human Performance, 22*, 375-403.

Kiesler, S., & Sproull, L. (1992). Group decision making and communication technology. *Organizational Behavior and Human Decision Processes, 52,* 96-123.

Leonardi, P., Jackson, M., & Marsh, N. (2004). The strategic use of 'distance' among virtual team members: A multidimensional communication model. In S.H. Godar & S.P. Ferris (Eds.), *Virtual and collaborative teams: Process, technologies & practice* (pp. 156-173). Hershey, PA: Idea Group Publishing.

Lipnack, J., & Stamps, J. (1997). *Virtual Teams: Reaching across space, time, and organizations with technology.* New York: John Wiley & Sons.

Majchrzak, A., Malhotra, A., Stamps, J., & Lipnack, J. (2004). Can absence make a team grow stronger? *Harvard Business Review, 82*(5), 131-137.

Majchrzak, A., Rice, R.E., King, N., Malhotra, A., & Ba, S. (2000). Technology adaptation: The case of a computer supported inter-organizational virtual team. *MIS Quarterly*, *24*, 569-600.

McCune, J.C. (1998). Telecommuting revisited. *Management Review, 87*, 10-16.

Napier, B.J., & Ferris, G.R. (1993). Distance in organizations. *Human Resource Management Review, 3*, 321-357.

Olson, G.M., & Olson, J.S. (2000). Distance matters. *Human Computer Interaction, 15,* 139-178.

Prashad, S. (2003, October 23). Building trust tricky for "virtual" teams. *Toronto Star*, K06.

Rosenfeld, L., Richman, J.M., & May, S.K. (2004). Information adequacy, job satisfaction and organizational culture in a dispersed-network organization. *Journal of Applied Communication Research, 32*, 28-54.

Scott, C.R., & Timmerman, C.E. (1999). Communication technology use and multiple workplace identifications among organizational teleworkers with varied degrees of virtuality. *IEEE Transactions on Professional Communication, 42,* 240-260.

Shockley-Zalabak, P. (2002). Protean places: Teams across time and space. *Journal of Applied Communication Research, 30*, 231-250.

Townsend, A.M., DeMarie, S.M., & Hendrickson, A.R. (1998). Virtual teams: Technology and the workplace of the future. *Academy of Management Executive, 12,* 17-29.

Upton, D.M., & McAfee, A. (1996). The real factory. *Harvard Business Review*, *74*(4), 123-133.

Van Aken, J.E., Hop, L., & Post, G.J.J. (1998). The virtual organization: A special mode of strong interorganizational cooperation. In M.A. Hitt, J.E. Ricart I Costa & R.D. Nixon (Eds.), *Managing strategically in an interconnected world* (pp. 301-320). Chichester, UK: John Wiley & Sons.

Warkentin, M.E., Sayeed, L., & Hightower, R. (1997). Virtual teams vs. face-to-face teams: An exploratory study of a Web-based conference system. *Decision Sciences, 28,* 975-996.

Wiesenfeld, B.M., Raghuram, S., & Garud, R. (1999). Communication patterns as determinants of identification in a virtual organization. *Organization Science, 10,* 777-790.

Yagil, D. (1998). Charismatic leadership and organizational hierarchy: Attribution of charisma to close and distant leaders. *Leadership Quarterly, 9*, 161-176.

Zaccaro, S.J., & Bader, P. (2003). E-leadership and the challenges of leading e-teams. *Organizational Dynamics, 31*, 377-387.

Zack, M.H. (1994). Electronic messaging and communication effectiveness in an ongoing work group. *Information and Management, 26,* 231-241.

KEY TERMS

Accessibility: An individual's perception that he/she can contact or reach his/her leader when so desired.

Dispersed/Distributed Teams: Teams separated by some degree of physical distance.

Distanced Leadership: Leadership of a team or organizational members that are separated by some degree of time and distance from their leader.

Identification: Identification is the process in which an individual comes to see an object (e.g., an individual, group, organization) as being definitive of oneself and forms a psychological connection with that object. Although scholars have offered a variety of conceptual definitions for identification, we view it as a communicative process, rooted in discourse and constituting a communicative expression of one's identity.

Social Presence: The perception of physical and/or psychological access to another. Social presence theory often focuses on the aspects of communication media that permit people to connect or "be present" with others and the theory sees some degree of social connectedness as crucial to work relationships.

The Dynamics of Virtual Teams

Norhayati Zakaria
Syracuse University, USA

Shafiz A. Mohd Yusof
Syracuse University, USA

INTRODUCTION

The world continues to be driven by the rapid development of information technology and globalization. Not surprisingly, the working environments that have been projected to grow the fastest are all related to the usage of computers, the Internet, and information systems. With globalization, many multinational corporations (MNCs) are increasingly employing virtual teams (VTs). It was reported that 137 million workers worldwide are involved in some form of remote electronic work (Solomon, 2001).

Some examples of companies that are already using virtual teams are Nortel Networks Corporation, which has 80,000 employees located in 150 countries; Price Waterhouse, which has 45,000 employees in 120 countries; and Deloitte & Touche LPP in New York, which has 90,000 employees in 130 countries (Geber, 1995; Solomon, 2001). The survey by Gartner Group Incorporation (Biggs, 2000) further estimated that 60% of the professional and management tasks at Global-2000 companies would be done via virtual teams by 2004.

In this article, we present the issues and challenges that are encountered by VTs in the same organization. In this respect, we will analyze VTs in the context of intraorganizational (within one organization) and not interorganizational (across different organizations). Our emphasis is on team members that have diverse cultural backgrounds and work in a global environment such as those in MNCs. Hence, a virtual team is defined as a group of members that collaborate and communicate primarily via computer-mediated communication (CMC) without any geographical boundaries, and the composition of the team members consists of people from different cultural backgrounds: for example, people from Motorola in Malaysia collaborating on a 12-week project with people from Japan and in the U.S. In essence, the project involves team members from three different countries—Malaysia, Japan, and the U.S.—yet all of them belong to the same organization: Motorola. Nonetheless, we do acknowledge that VTs can also include team members with similar cultural backgrounds, or team members that belong to different or multiple organizations, but both aspects are not the emphasis in our article.

BACKGROUND

The conceptual and empirical research conducted on the topic of VTs has also increased tremendously over the past years when electronic-mediated technologies became more ubiquitous (Belbin, 1981; Geber, 1995; Jarvenpaa & Leidner, 1999; Kostner, 1994; Townsend, DeMarie, & Hendrickson, 1996). Many researchers began to look at many different issues faced by virtual teams. Some of the latest works that concern VTs are studies that look at personality types and interaction styles that affect the communication of VTs as compared to conventional team performances (Potter & Balthazard, 2002), the use of technology in global virtual collaboration (Qureshi & Zigurs, 2001), the effects of the temporal coordination mechanism on conflict management, the behavior of VTs supported by an asynchronous communication technology (Montoya-Weiss, Massey, & Song, in press), radical innovations without collocation (Malhotra, Majchrzak, Carmen, & Lott, 2001), the understanding of the best practices of VTs (Lurey & Raisinghani, 2001), sharing and reusing knowledge between team members in other organizations, and virtual relationships (Majchrzak, Rice, King, Malhotra, & Ba, 2000). Problems stemming from intercultural communication, trust, leadership, and training are all crucial to understanding in light of VTs.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

According to Maznevski and Chudoba (2000), out of the 41 studies conducted on technology-supported distributed teams from 1990 to 1998 published in 11 major journals, only a small number of research works were conducted to understand how cultural boundaries affect the context in which the communication takes place and the communication process itself (e.g., Turoff, Hiltz, Bahgat, & Rana, 1993). Furthermore, for internationally distributed teams, only two studies were conducted to understand the role of trust in global teams that never met face to face (Jarvenpaa, Knoll, & Leidner, 1998; Jarvenpaa & Leidner, 1999). Research on multinational teams is far more limited than research on distributed teams, with most of it focusing on the effectiveness of team performance of a heterogeneous group and a homogenous group (Maznevski & Chudoba, 2000).

Teams are often viewed as an important means to enhance an organization's creative and problem-solving capabilities (Jarvenpaa, Ives, & Pearlson, 1996; Zachary, 1998). Maznevski and Chudoba (2000) define global virtual teams as groups that (a) are identified by their organization(s) and members as a team, (b) are responsible for making and/or implementing decisions important to the organization's global strategy, (c) use technology-supported communication substantially more than face-to-face communication, and (d) work and live in different countries. A virtual team is also defined as "a temporary, culturally diverse, geographically dispersed, electronically communicating work group" (Jarvenpaa & Leidner, 1999, p. 792). The notion of temporary in the definition describes team members that may have never worked together before and who may not have expected to work together again as a group (Jarvenpaa & Ives, 1994; Lipnack & Stamps, 1997). A virtual team is considered global when the members' backgrounds are culturally diverse and they are able to think and work with the diversity of the global environment (DeSanctis & Poole, 1997; Jackson et al., 1995). Finally, the team members use computer-mediated communication technology such as groupware that allows members to engage in collaborative work despite the separation of time and space.

ISSUES AND CHALLENGES FACED BY VIRTUAL TEAMS

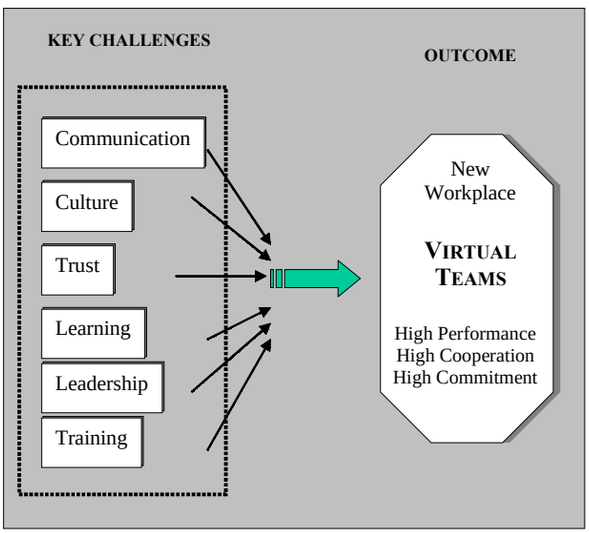
There are some challenges and issues facing the VTs that we need to address. First of all, the emergence of VTs implies the extensive use of electronic forms of communication such as e-mail, videoconferencing, online discussion forums, the Internet, and so forth. The complexity in communicating over time, distance, and space causes the MNCs unique problems that are not easy to solve. Although distance and speed can be considered the most desirable advantages to VTs, there are other essential aspects that would create some problems and challenges such as the lack of expressive (nonverbal) behavioral cues as well as contextual cues (Sproull & Kiesler, 1986).

Practical experiences and research findings have shown that when VTs are not managed properly, they can be less effective than traditional teams. In addition, Warekentin, Sayeed, and Hightower (1997) reported on studies that have shown that teams that rely entirely on virtual communication by substituting face-to-face communication would resist using this form of communication. However, living in a global village, many organizations realize and take into consideration that VTs are their best resource of human assets. Townsend, Hendrickson, and DeMarie (2002) postulate that the reality of the virtual workplace accentuates the need for change in management, particularly when organizations need to understand some of the inherent challenges that people face at the transition from traditional teamwork to a virtual one. The following discussions pertain to the challenges and issues facing VTs (as illustrated in Figure 1).

Understanding Diverse Cultural Values: Organizational and National

It is essential to note that cultural issues are often overlooked by global managers simply because it is intricate and difficult to understand. Yet this issue is crucial if global managers need to fully understand, be aware, and be sensitive to the managerial outcomes and implications of working in a virtual environment. In addition, global managers also need to develop effective strategies and tactics on how to best overcome the culturally conflicting issues given

Figure 1. A framework for understanding effective virtual teams



that the individual team members comprise multicultural backgrounds.

There are two forms of cultural values that MNCs need to manage: organizational and national cultures. Since this article focuses on the intraorganizational VTs, the organizational culture therefore deals with the complex internal needs of an organization. It involves the intricate matter of culture, which includes both the physical environment and the internal supporting environment in which technology will be used. Organizational culture is like a glue that influences, motivates, and directs people in selecting the direction or goal they are heading in an organization. It is essential to note that communicating electronically is a challenging phenomenon to many teams at the workplace. Team members would feel anxious and uncertain unless the distant members clearly understand the communication that takes place and the activities and tasks that each of them are involved in. Furthermore, some people regard that this virtual work arrangement is not following the norm of working, which makes it even harder for them to adapt. Hence, this presents more resistance to adopt technology.

In order to introduce effective ways of working by means of technology, it is important for top management to support the use of the technology. The use of technology can be in a variety of forms, ranging from telephones, faxes, teleconferences, e-mail, videoconferences, collaborative design tools, and knowledge-management systems (Gibson & Cohen, 2003). Organizations, therefore, must have the right culture for the new structures, processes, procedures, and innovation. Any form of change can create a feeling of uncertainty and anxiety in people, be it organizational or technological. Resistance will surface unless the culture is receptive to changes and people are ready and able to accept new ideas. It is difficult to make any changes, especially if it involves a major or abrupt transformation, unless people are motivated to keep changing for improvements.

Apart from organizational culture, national culture differences can cause the same challenges. National culture is defined as a way of life as it characterizes a certain set of shared behaviors, thinking, beliefs, and attitudes of the people. According to Hofstede (1980), culture is defined as a collective mental programming that conditions people's values and perceptions. Fundamentally, culture is an idea about the world that influences how people think, feel, and act. Those assumptions arise from the shared mental models and experiences of a group of people. Because different groups speak different languages and have different experiences, they construct different visions of the world. When people from different cultures come in contact with one another (which they inevitably do in a teamwork environment), those distinctive visions of the world and ways of doing things may collide or combine over time, or coexist disharmoniously. Thus, when culture clashes persist over time, this could potentially create chaos and an unpleasant working environment for people to work in. Each of these potential chaotic outcomes happens, though, at least partly because of miscommunication and misinterpretation, for example, of messages sent through e-mail.

Moreover, people need to effectively communicate before they can collaborate. As a team, they need to work hand in hand regardless of whether they are working virtually or physically. The wide-ranging forms of technology-mediated communication can pose a challenge to the team members who have different cultural values when they collaborate and communicate. It is useful to note that technology or medium preferences can vary based on the richness of the medium: from the high in richness, for example, videoconferencing, which is effective for

immediate feedback and personalization, to the low-in-richness (leanest) forms of technology, such as e-mail, which is less effective for immediate feedback and its lack of nonverbal cues (Daft & Lengel, 1986). Collectivistic or high-context people like Asians and Arabs desire face-to-face communication or a rich form of technology medium better because they rely heavily on nonverbal cues such as body language, tone of voice, facial expressions, and gestures to understand the communication and information shared. The absence of nonverbal cues makes it harder for people to interpret the subtle meanings that are embedded in the cues. Collectivistic cultures also prefer speech that uses indirect, ambiguous, and subtle language (Hall, 1976; Hofstede, 1980). As a result, high-context people may prefer to use a technology medium such as videoconferencing to collaborate and communicate among team members.

On the contrary, individualistic and low-context people like Americans and British may be frustrated and confused when they try to interpret and understand the meaning of indirect requests because they use a more direct, detailed, and explicit language when communicating. They also prefer content of the message that focuses on words and verbal language as compared to context that focuses on nonverbal elements. Hence, low-in-richness technology media such as e-mail may suffice for them since the medium is heavily based on text and words when sharing information and communicating.

Building Swift Trust Among Virtual Team Members

According to Das and Teng (1998), intercultural communication competence is an important antecedent to trust in the context of team alliances. Teams are able to know the level of trustworthiness of their members through effective communication. Intercultural communication is challenging in this case because it involves many different styles and patterns (Chen, 2001; Gudykunst & Ting-Toomey, 1996). According to Sitaram (as cited in Novinger, 2001, p. 4), "[b]ecause different cultures often demand very different behaviors, intercultural communication is more complex than communication between persons of the same culture." In order to function effectively and to increase confidence and security in cross-cultural relationships, VTs require trust (Earley, 1994). Trust is the critical enabling condition and the glue of the global workspace (Duarte & Snyder, 1999; Gibson & Cohen, 2003; Jarvenpaa & Leidner, 1999). Since a virtual team is a phenomenon that is based on temporal and ad hoc relationships, people need to form trust in a short time, a notion called swift trust. In a virtual environment, the problem of building swift trust is exacerbated given the cultural differences of the team members and the type of technology used for communication. A need to explore and understand the trust phenomenon is thus fundamental given the avid use of VTs in MNCs.

The concept of teams varies across cultures and organizations, and how teams are perceived will differ based on the organizational and national cultural attributes of its members (Gibson & Zellmer-Bruhn, 2001). It is also critical to note that individuals from different national cultures vary in terms of their group behaviors and communication styles (Gudykunst, 1997), which in turn impacts the formation of swift trust. Empirical studies that explain or even understand the impact of cultural diversity on communicative behaviors for effective trust building and information sharing in VTs are largely deficient. As a result, the need to understand how to improve team effectiveness and what facilitates trust and information sharing between cross-cultural teams is paramount (Duarte & Snyder, 1999; Kostner, 1994; Lipnack & Stamps, 1997). According to Jarvenpaa and Leidner (1999), trust needs to be formed in the very early stage of a relationship, so the first interactions are strategic. They went on to emphasize that "virtual team members should be very careful about what they say in their first messages" (p. 2). Jarvenpaa and Leidner further added that in the virtual world, the old adage of "you never get a second chance to make a first impression" reflects the high need for effective intercultural communication competence.

Trust is considered as a key lubricant for cross-cultural relationships. Johnson and Cullen (2002, p. 335) suggested that "in exchange relationships, where one party's outcome depends on the behavioral and intent of the exchange partner, trust is particularly crucial. Without trust, the objectives and outcomes of the exchange are in constant and chronic jeopardy." Trust also provides a crucial condition for informa-

tion-sharing behaviors. The main concern for this psychological process of information sharing is that people attempt to control the flow of intimate, personal, or private information. Hence, once trust is achieved, team members are more willing to disclose and share information via electronic-mediated technology.

Learning to Lead and Train Virtual Teams

Within a business environment, learning is a conscious attempt on the part of organizations to improve productivity, effectiveness, and innovativeness in uncertain economic and technological market conditions. The greater the uncertainties in the virtual work environment, the higher the need for learning to take place among the team members. Learning can be challenging for team members that have never experienced working in a virtual environment. Not only do they need to learn how to use the technology for effective communication, but they also need to learn about the cultural backgrounds of their remote-located team members. Learning to develop instant or swift rapport and trust is another challenging issue. These issues can be overwhelming and stressful.

Nonetheless, learning is successful when people are ready and willing to learn; that is, when people are motivated and curious to know something. Sometimes people's readiness to learn comes with time and experience. In this case, a leadership role can be a major factor in motivating one's learning process. If a desired change in behavior is imperative, the leader may need to supervise directly to ensure that the desired behavior occurs. But, in a virtual environment, empowerment is the key to self-learning. Hence, absolute support from leaders such as top management can result in learning to occur naturally. There is no doubt that leadership plays an important role in virtual workplaces. This role also assumes a different responsibility since team members are empowered. Leaders should recognize that leading and managing VTs are two key strategies that need to be carried out jointly for organizations to succeed.

The last challenge arises from the issue of training the VTs. There are two types of training that are critical: training teams for intercultural competence and training for technological competence. When one has an ability to interact across different cultural contexts and become aware of one's own and others' cultural conditioning, one is known as having intercultural communication competence (Gudykunst & Ting-Toomey, 1996). Intercultural communication involves an exchange of meaning in which the process of communicating is more dynamic, multifaceted, and complex. As such, cultural conditioning will affect the evaluation of experiences as well as the means by which information and knowledge is shared, conveyed, and learned. In a culturally diverse environment, the transmission of information does not often, however, ensure understanding and learning. Typically we view the transmission of information from sender to receiver as a one-way process where the active participant is the sender while the receiver remains an inactive recipient. From a multicultural perspective, the challenge stems from the transmission of information that involves many people at one time from different locations and cultural backgrounds.

Even though computer-mediated technology has become pervasive in today's workplace, there is growing evidence of unrealized or less-than-expected productivity gains due to poor technology acceptance and use by users or the employees. Hence, this circumstance entails MNCs to train people for technological competence. Naturally, people always resist any new ways of doing things. Having the right attitude and perception of using technology is critical. In a review of computer attitude training, Dupagne and Krendl (1992) and Liu, Reed, and Philips (1992) showed that people who have no prior experience exhibited high anxiety toward using computers, while people who have had computer training are more likely to show positive attitudes toward computer use.

In essence, we came up with the following implications to illustrate some of the key challenges that VTs encounter in the new workplace. Once the issues and challenges are overcome, the result leads to enhanced performance, cooperation, and commitment from the VTs.

IMPLICATIONS FOR MULTINATIONAL CORPORATIONS

Due to the heavy reliance on CMCs for this global and virtual work context, MNCs need to employ strategies

on how to effectively motivate, train, and lead VTs to collaborate. The current and future trends for MNCs are increasing involvement in high-intensity collaboration among the virtual team members. As the pace of change is rapid, this condition necessitates MNCs to renew themselves and constantly adopt innovative ideas to keep up with the challenges. With remote or virtual workplaces, team members are being empowered, which means that each team is accountable for success and is a source of creative thinking. All workers need to become part of an organization-wide collaboration process with standard operating processes and procedures.

Strategies like open communication where members provide honest feedback, accept constructive criticism, and address issues head-on need to be employed. Trust requires leaders. In the most ideal situation, the units that are in good trust-based organizations hardly have to be managed. Managing VTs thus takes on a new perspective all together. What MNCs need is a distributed form of leadership. Because the emergence of this type of leadership in VTs is still limited, MNCs need to continuously promote and build leadership skills by establishing a strong organizational culture—a culture that can promote learning, trust, and teamwork values as its key monitoring mechanisms (Zakaria, Amelinckx, & Wilemon, 2004).

In a similar vein, it is useful to note that MNCs need to incorporate complementary sets of leadership skills that are technical as well as cross-cultural in nature. One cannot do without the other. If a person only knows how to use a technology but lacks the capability to understand human aspects such as cross-cultural differences, then the teamwork will not be successful. Therefore, training needs to be balanced in order to accommodate and instill this new set of balanced competencies. The future trend for selecting, recruiting, motivating, and training VTs should be an increase in the need for leaders and teamwork that fully understand human behavior. With this value, they could potentially engender cooperation and collaboration, and stimulate creativity and innovation. As management hierarchy becomes more flattened, global leaders will depend more than ever on these skills to create this high-commitment, high-performance, high-cooperation workplace.

FUTURE RESEARCH TRENDS AND CONCLUSION

With the demands of globalization and the integration of culturally diverse virtual teams, cross-cultural and technological management is becoming a more challenging task. The relationship between VTs and CMCs will become increasingly important as organizations increasingly rely on virtual teams to carry out and implement important projects. Deploying CMCs intensifies the challenges of global management, whether it is team based or not. For example, CMCs can shape the way people perform their tasks in organizations as much as they can impact the way people communicate and collaborate globally. Without a doubt, we are focusing on new ways of working across borders. CMCs are not just simple tools; instead, they need to be integrated and aligned with team design, behavior, and the processes of collaborating and communicating. While CMC usage is essential in the communication and knowledge-sharing processes for geographically dispersed employees, computer-facilitated communication technologies are only as effective as those using them.

Based on several reviews of the literatures, there are many future research opportunities that can help advance the knowledge of VTs and their use of CMCs. Zakaria et al. (2004) suggested that research on the following areas of inquiry could offer many useful insights into improving the effectiveness of these increasingly important teams.

- How important is a sense of team membership in VTs? If it is important, what are the most effective means for developing VTs' identification? How might the use of CMCs assist in this process?
- What approaches can be used to resolve conflicts in VTs? Are traditional conflict-management techniques useful in a VT environment? How might the use of CMCs increase and/or decrease the amount of conflict experienced in a VT?
- How does leadership emerge and how is it exercised in VTs? What factors are particularly important in gaining and maintaining support? How might CMCs assist in the exercise of the leadership function?

- What role does trust play in VTs? If swift trust is an important factor in facilitating virtual teamwork, what function might CMCs play in creating and maintaining such trust?
- What role can CMCs play in helping to overcome the cultural barriers that can hinder effective VT performance?
- How can VT members most effectively transfer ideas, plans, resource needs, and performance objectives to their respective organizations via CMCs?
- Within an organization, how can the learning from one VT be captured, stored, and retrieved by a newly appointed leader in VTs? What role can CMCs play in the VT learning and dissemination process?
- Can useful online training programs be designed for VTs to develop technological and intercultural communication competence? If so, how can these programs be administered?

We conclude that both the human and technology aspects need to be managed so that a virtual team with high performance, high commitment, and high cooperation can be created. Therefore, MNCs need to select and recruit the right people to work in the virtual workplace—people with open minds and flexible attitudes, as well as people that are willing to collaborate and work in teams. It is indispensable for MNCs to incorporate an organizational culture that acclimatizes to factors such effective leadership, efficient use of CMCs, and appropriate rewards and incentives for VTs to perform successfully. Last, national culture should not be viewed as a barrier for effective collaboration; instead, the differences should be viewed as a synergy to create an innovative workplace.

REFERENCES

Belbin, R. M. (1981). *Management teams: Why they succeed or fail.* Oxford, UK: Butterworth-Heinemann.

Biggs, M. (2000). Enterprise toolbox: Assessing risks today will leave corporate leaders well prepared for the future of work. *InfoWorld, 22*(39), 100-101.

Chen, G. M. (2001). Toward transcultural understanding: A harmony theory of Chinese communication. In V. H. Milhouse, M. K. Asante, & P. O. Nwosu (Eds.), *Transcultural realities interdisciplinary perspectives on cross-cultural relations* (pp. 55-70). Thousand Oaks, CA: SAGE.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571.

Das, T. K., & Teng, B. S. (1998). Between trust and control: Developing confidence in partner cooperation in alliances. *Academy of Management Review, 23*(3), 491-512.

Duarte, D. L., & Snyder, N. T. (1999). *Mastering virtual teams: Strategies, tools, and techniques that succeed*. San Francisco: Jossey-Bass Publishers.

Dupagne, M., & Krendl, K. A. (1992). Teachers' attitude toward computers: A review of the literature. *Journal of Research on Computing in Education, 24*(3), 420-429.

Earley, P. C. (1994). Self or group? Cultural effects of training on self-efficiency and performance. *Administrative Science Quarterly, 39*, 89-117.

Geber, B. (1995). Virtual teams. *Training, 32*(4), 36-41.

Gibson, C. B., & Cohen, S. G. (2003). *Virtual teams that work: Creating condition for virtual teams effectiveness.* San Francisco: Jossey-Bass.

Gibson, C. B., & Zellmer-Bruhn, M. E. (2001). Metaphors and meaning: An intercultural analysis of the concept of teamwork. *Administrative Science Quarterly, 46*, 274-303.

Gudykunst, W. B. (1997). Cultural variability in communication. *Communication Research, 24*(4), 327-348.

Gudykunst, W. B., & Ting-Toomey, S. (1986). Communication in personal relationships across cultures: An introduction. In W.B. Gudykunst, S. Ting-Toomey & T. Nishida (Eds.), *Communication in personal relationships across cultures* (pp. 3-16). Thousand Oaks, CA: SAGE.

Hall, E. T. (1976). *Beyond culture*. Garden City, NJ: Anchor Books/Doubleday.

Hofstede, G. (1980). *Culture's consequences: International differences in work related values.* Beverly Hills, CA: Sage.

Jackson, S. E., May, K. E., & Whitney, K. (1995). Understanding the dynamics of diversity in decision-making teams. In R.A. Guzzo & E. Salas (Eds.), *Team effectiveness and decision making in organizations* (pp. 7-261). San Francisco: Jossey-Bass.

Jarvenpaa, S. L., & Ives, B. (1994). Transitions in teamwork in new organizational forms. *Advances in Group Processes, 14,* 157-176.

Jarvenpaa, S. L., & Leidner, D. E. (1999). Communication and trust in global virtual teams. *Organization Science, 10*(6), 791-815.

Jarvenpaa, S. L., Ives, B., & Pearlson, K. (1996). Global customer service for the computer and communication industry. In P. C. Palvia, S. C. Palvia, & E. M. Roche (Eds.), *Global information technology & system management: Key issues and trends*. Westford, MA: Ivy Publishing.

Jarvenpaa, S. L., Knoll, K., & Leidner, D. (1998). Is anybody out there? Antecedents of trust in global teams. *Journal of Management Information Systems, 14*(4), 29-64.

Johnson, J. L., & Cullen, J. B. (2002). Trust in cross-cultural relationships. In M. J. Cannon & K. L. Newman (Eds.), *The Blackwell handbook of cross-cultural management* (pp. 335-360). Malden, MA: Blackwell Publishing.

Kostner, J. (1994). *Virtual leadership: Secrets from the round table for the multi-site manager*. New York: Warner.

Lipnack, J. P., & Stamps, J. S. (1997). *Virtual teams: Reaching across space, time and organizations with technology*. New York: John Wiley & Sons.

Liu, M., Reed, W. M., & Philips, P. D. (1992). Teacher education students and computers: Gender, major, prior computer experience, occurrence, and anxiety. *Journal of Research on Computing in Education, 24*(4), 457-467.

Lurey, J. S., & Raisinghani, M. S. (2001). An empirical study of best practices in virtual teams. *Information & Management, 38*(8), 523-544.

Majchrzak, A., Rice, R. E., King, N., Malhotra, A., & Ba, S. (2000). Technology adaptation: The case of computer-supported interorganizational virtual team. *MIS Quarterly, 24*(4), 569-600.

Malhotra, A., Majchrzak, A., Carman, R., & Lott, V. (2001). Radical innovation without collocation: A case study at Boeing-Rocketdyne. *MIS Quarterly, 25*(2), 229-249.

Maznevski, M. L., & Chudoba, K. M. (2000). Bridging space over time: Global virtual team dynamics and effectiveness. *Organization Science*, *11*(5), 473-492.

Montoya-Weiss, M., Massey, A. P., & Song, M. (in press). Getting it together: Temporal coordination and conflict management in global virtual teams. *Academy of Management Journal, 44*(6), 1251-1262.

Novinger, T. (2001). *Intercultural communication*. Austin, TX: University of Texas Press.

Potter, R. E., & Balthazard, P. A. (2002). Understanding human interactions and performance in the virtual team. *Journal of Information Technology Theory & Application, 4*, 1-23.

Qureshi, S., & Zigurs, I. (2001). Paradoxes and prerogatives in global virtual collaboration. *Communications of ACM, 44*(12), 85-88.

Schein, E. (1992). *Organizational culture and leadership* (2nd ed.). San Francisco: Jossey-Bass.

Solomon, C. (2001). Managing virtual teams. *Workforce, 80*(6), 60.

Sproull, L. & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science, 32,* 1492-1512.

Townsend, A. M., Hendrickson, A. R. & DeMarie, S. M. (2002). Meeting the virtual work imperative: How collaborative technologies can facilitate face-to-face meetings without the need to even be in the same room. *Communications of the ACM, 45*(1), 23-26.

Turoff, M., Hiltz, S. R., Bahgat, A. N. F., & Rana, A. R. (1993). Distributed group support system. *MIS Quarterly, 17*(4), 399-417.

Tyler, E. B. (1871). *Origins of culture*. New York: Harper & Row.

Warkentin, M. E., Sayeed, L., & Hightower, R. (1997). Virtual teams versus face-to-face teams: An exploratory study of a Web-based conference system. *Decision Sciences, 28*(4), 975-996.

Zachary, G. P. (1998). The rage for global teams. *Technology Review, 101*(4), 33.

Zakaria, N., Amelinckx, A., & Wilemon, D. (2004). Working together apart? Building a knowledge-sharing culture for global virtual teams. *Creativity and Innovation Management, 13*(1), 15-29.

KEY TERMS

Computer-Mediated Communication (CMC): It describes all media that are involved in the dynamic transfer and storage of data (analogue and digital) across established networks. The technology includes the World Wide Web, e-mail, telephones, fibre optics, and satellites.

High-Context Culture: This is a situation when one looks for information in the physical context or internalized in the person while very little is a coded, explicit, or transmitted as part of the message.

Intercultural Communication Competence: This is defined as the ability to effectively and appropriately execute communication behaviors to elicit a desired response in a specific environment.

Low-Context Culture: This is a situation where the mass of the information is vested in the explicit code. Thus, in a low-context culture, senders assume little or no shared knowledge with receivers.

National Culture: This is a collective mental programming that conditions people's values and perceptions. Culture can also be defined as "that complex whole that includes knowledge, belief, art, morals, law, custom and any other capabilities and habits acquired by man as a member of a society" (Tyler, 1871, p. 1).

Organizational Culture: Organizational culture is a pattern of shared basic assumptions that the group learned as it solved its problems of external adaptation and internal integration that has worked well enough to be considered valid. Therefore, it is taught to new members as the correct way to perceive, think, and feel in relation to those problems.

Values: These are defined as desirable states, objects, goals, or behaviors transcending specific situations and applied as normative standards to judge and to choose among alternative modes of behavior.

Virtual Team: According to Gibson and Cohen (2003), a virtual team must have the following attributes: (a) It is a functioning team in which the individuals are interdependent in their tasks, share responsibility for outcomes, and are intact as a social unit; (b) members are geographically dispersed; and (c) technology-mediated communication is relied upon to accomplish the task. A virtual team can have culturally diverse team members or culturally similar team members that are located in a physically dispersed environment and not collocated. They also normally work in a temporary collaborating work group.

E-Commerce and Usability

Shawren Singh
University of South Africa, South Africa

ELECTRONIC COMMERCE

The term "Electronic Commerce" (EC) conjures various interpretations. Figure 1.1 shows some of the different types of EC, of which there are many, such as Business-to-Business (B2B); Business-to-Consumer (B2C); Consumer-to-Business (C2B); Consumer-to-Consumer (C2C); People-to-People (P2P); non-business EC; intrabusiness (organisational) EC; business-to-employees (B2E); government-to-citizen (G2C); exchange-to-exchange (E2E); collaborative commerce; ultimate commerce (u-commerce) and mobile commerce (m-commerce).

A certain basic infrastructure is required for any type of EC to function efficiently. Key components of this infrastructure are networks, Web servers, Web servers' support and software, electronic catalog, Web page design and construction software, transactional software and Internet access components (Turban, Rainer & Potter, 2001).

The infrastructure on which the EC application is built will affect the users' experience of that application. It is generally accepted that any EC that does not provide the user with such experience will not thrive (Brandt, 1999). The traditional approaches of enticing a purchase in brick-and-mortar commerce, such as atmosphere, placement of goods, lightning and so forth, cannot be applied to online commerce. Nielsen (1999) contends that a "bad" user interface is one of the reasons for EC failure. Interaction and participation are the emotional hooks for EC, and the developers of EC sites should bear this in mind.

DEVELOPMENT OF EC APPLICATION

As companies realised that their EC ventures were not as successful as they had anticipated and were prone to failure, they began to investigate alternate development strategies to deal with this rapidly changing environment. One such approach that has won favour amongst Web application developers is agile development methodologies (ADMs). ADMs do not have prescriptive processes and do not define detailed procedures on how to create a given type of model. Instead, they provide advice on how to be effective as a modeller. As opposed to the traditional development approaches, ADMs are not hard and fast. ADMs can also be seen as a craft and not a science (Ambler, 2003).

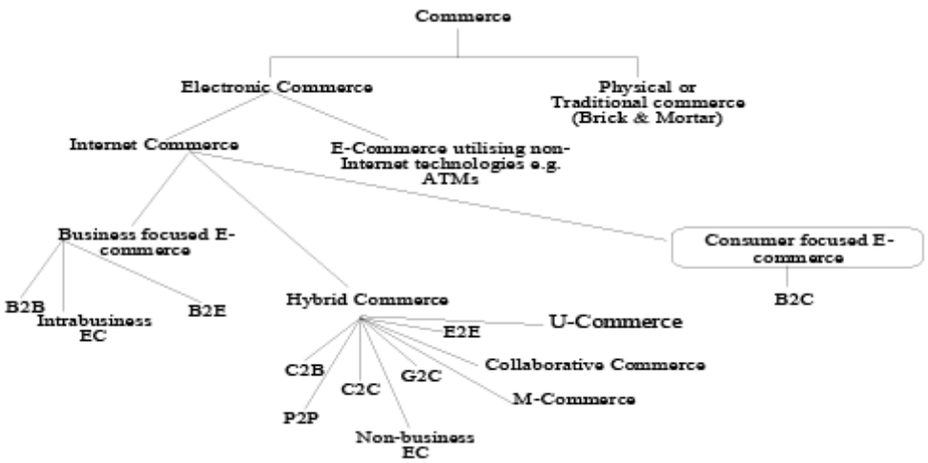
With regard to the World Wide Web (WWW), Pressman (2000, p. 8) states: "What worries me is that this major new technology has become a breeding ground for important Web applications that are hacked in much the same way as important application software was hacked a few generations back—in the 1960s and 1970s."

Pressman (2000) goes on to say that the current basic Web development philosophy is that Web applications must be developed within days or weeks. The argument is that time frames do not allow for anything but a rush to the finish line. Web applications are constantly evolving. The argument, then, is why spend time on specifying what is needed and designing how to build it when everything will change anyway? Web applications are inherently different from other application software. The argument is that the content (text, graphics, images, audio and video, for example) is inextricably integrated with procedural processing.

Pressman believes that people who use Web applications are more tolerant of errors. The argument is that users really want exciting Web sites that are up and running in days. He argues that it is almost impossible to know what Web applications users really want, because the demographics of Web visitors are so hard to predict. We believe that as Web applications are becoming an integral part of life, users' fault tolerance is becoming much lower. The people who build Web applications are different. Web developers are free thinkers, who certainly would feel unduly constrained by the old ways. In fact, talk of a disciplined approach, other than "build it, test it to death (if time permits), and then put it online," usually results in grimaces all around. The development of

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Types of commerce (adapted from Chan, Lee, Dillon and Chang (2001)



applications for the WWW, therefore, has its own set of unique problems. No current theory adequately addresses how to effectively create Web sites for online selling.

USABILITY

Usability refers to how usable a system is from a user's point of view. Usability concerns are not about presentation, but a whole gamut of exchanges that form human-computer interaction. Below are some issues that need to be considered to improve usability.

COGNITION, PERCEPTION AND PHYSIOLOGY

We can divide human (user) resources into three categories (Kotze, 2000): *Perception*, which refers to the way in which humans detect information in their environment; c*ognition,* which refers to the way in which humans process that information; and *physiology,* which refers to the way in which humans move and interact with physical objects in their environment.

A vital foundation for interactive-system designers is an understanding of the cognitive, perceptual and physiological abilities of the user. The human ability to interpret sensory input rapidly and to initiate complex actions makes modern computer systems possible. *Perception* involves the use of our senses to detect information (Kotze & Johnson, 2001). In computerized systems, this mainly involves using senses to detect audio output, senses to detect visual output and tactile feedback. This is affected by many factors, such as change in output (loudness/size); maximum and minimum detectable levels; field of perception (can the user see the display?); fatigue; Circadian rhythms (biological rhythms); problems with background noise; and so forth. *Cognition* involves various cognitive processes, including (Kotze & Johnson, 2001) short-term memory; long-tem memory and learning; problem solving; decision making; attention and scope of concern; search and scanning; time perception; perceptual or mental load; anxiety; and fear.

When we operate a system, we gradually move from general knowledge to rules and then to skills. Users with greater expertise will be able to enter the process at a higher level. Ideally, we all want to work at the highest skill level. We do not want to spend time thinking about the use of previous systems or sifting through our general knowledge. The more we work at the knowledge and rule level, the more uncertain we are about things. Users do not want to be forced to make guesses. Guessing introduces inefficiency and can consume a great deal of time in "repair" tasks when things go wrong; for instance, if we delete a file by accident.

The more we have to think about using the interface, the fewer cognitive and perceptual resources we will have at our disposal for our main task.

Physiology involves the study of human anatomy (Kotze & Johnson, 2001). It might seem strange to include a discussion on physiology when discussing user interface design, but it can have a critical impact upon the design of a successful system. As a minimum requirement, users must be able to "view" the display, reach the input devices and so forth. A number of factors may intervene to restrict/prevent users from achieving this. Do not make interface objects so small that they cannot be selected by a user in a hurry, carrying a stack of books. Do not make disastrous options so easy to select that they can be started by accident.

It is important to note that interfaces often tend to reflect the assumptions that their designers make about the physiological characteristics of their users. Buttons are designed so that an "average" user can easily select them with a mouse, touchpad or trackball. Unfortunately, there is no such thing as an average user. Some users have the physiological capacity to make fine-grained selections, but others do not. Even if systems are unaffected by these issues, it is good to remember that workplace pressures of time and concentration may reduce the physiological ability of users.

The flexibility of computer software makes it possible for designers to provide special services for users who have disabilities such as visual, hearing and mobility impairments. Enlarging portions of a display or converting displays to Braille or voice output can be done with hardware and software supplied by many vendors. Text-to-speech conversion can help blind users to receive electronic mail or read text files, and speech-recognition devices permit voice-controlled operation of some software. Graphical user interfaces were a setback for vision-impaired users, but technological innovations facilitate conversion of spatial information into non-visual modes.

Users with hearing impairments can often use computers with only simple changes (conversion of tones to visual signals is often easy to accomplish), and can benefit from office environments that make heavy use of electronic mail and facsimile transmission.

Special input devices for users with physical disabilities will depend on the user's specific impairment. Devices available include speech recognition, eye-gaze control, head-mounted optical mice and so forth.

Designers can benefit by planning early on to accommodate users who have disabilities, since substantial improvements can be made at low or no cost.

CULTURAL AND PERSONALITY DIFFERENCES

Some people dislike computers or are made anxious by them; others are attracted to or eager to use them (Kotze, 2000). Often, members of these divergent groups disapprove or are suspicious of members of the other community. Even people who enjoy using computers may have different preferences for interaction styles, pace of interaction, graphics vs. tabular presentations, dense vs. sparse data presentation, step-by-step work vs. all-at-once work and so forth. These differences are important. A clear understanding of personality and cognitive styles can be helpful in designing systems for a specific community of users.

Another perspective on individual differences has to do with cultural, ethnic, racial or linguistic background (Kotze, 2000). It seems obvious that users who were raised learning to read Japanese or Chinese will scan a screen differently from users who were raised to read English or Afrikaans. Users from cultures that have a more reflective style or respect for ancestral traditions may prefer interfaces different from those chosen by users from cultures that are more action-oriented or novelty-based.

The term "culture" is often wrongly associated with national boundaries. Culture should rather be defined as behavior typical of a group or class of people. Culture is conceptualized as a system of meaning that underlies routine and behavior in everyday working life. Culture includes race and ethnicity as well as other variables and is manifested in customary behaviors, assumptions and values, patterns of thinking and communicative style.

As software producers expand their markets by introducing their products in other countries, they face a host of new interface considerations (Kotze, 2000). Little is known about computer users from different cultures, but designers are regularly called

on to create designs for other languages and cultures. The growth of a worldwide computer market means that designers must prepare for internationalization. Software architectures that facilitate customization of local versions of user interfaces should be emphasized. The simplest problem is the accurate translation of their product to the target language. For example, all text (instructions, help, error messages, labels) might be stored in files, so that versions in other languages could be generated with little or no programming. Hardware concerns include character sets, keyboards and special input devices. Other problems include sensitivity to cultural issues, such as the use of images and color. User interface design concerns for internationalization are long and full of pitfalls. Whereas early designers were often excused for cultural and linguistic slips, the current highly competitive atmosphere means that more effective localization will often produce a strong advantage. Nowhere else is this more true than in the e-commerce environment.

PRINCIPLES AND GUIDELINES TO SUPPORT USABILITY

Dix (1998), for example, put forward principles to support usability in three categories: *Learnability*, *flexibility* and *robustness. Learnability* refers to the ease with which new users can begin effective interaction and then attain a maximal level of performance. Usability principles related to learnability include predictability, synthesizability, familiarity, generalizability and consistency. *Flexibility* refers to the multiplicity of ways in which the user and the system exchange information. A user is engaged with a computer to achieve some set of goals in the work or task domain. Usability principles related to flexibility include dialogue initiative, multi-threading, task migratability, substitutivity and customisability. *Robustness* refers to the level of support given to the user in determining successful achievement and assessment of goals. Usability principles related to robustness include observability, recoverability, responsiveness and task conformance.

Shneiderman (1998) also focused on this aspect. He advocates three groups of principles when he discusses user-centered design. Many of these overlap with the principles proposed by Dix et al. (1998). Shneiderman's (1998) principles include recognition of diversity, use of the eight golden rules of interface design and prevention of errors.

All applications require user interfaces, the design of which is not a trivial matter. The same is true for e-commerce and any other Web-based applications. Shneiderman (1998) states that within the ocean (WWW) of information "there are also lifeboat Web pages offering design principles, but often the style parallels the early user-interface writings of the 1970s." The problem of early user interfaces, ignoring the abilities and preferences of the users, is therefore still present.

Nielsen (1996) focuses specifically on the user interface and the usability of Web applications, and identifies 10 common mistakes that Web page authors make: Using frames; gratuitous use of bleeding-edge technology; complex user resource locations (URLs); long scrolling pages; lack of navigation support; nonstandard link colors; scrolling text, marquees and constantly running animations; orphan pages; long download times; and outdated information.

CONCLUSION

Following usability principles would not necessarily guarantee a successful interactive system, but would go a long way towards preventing major disasters or failures of e-commerce activities.

REFERENCES

Ambler, W.S. (2003). What is Agile Modeling (AM)? Retrieved October 20, 2003, from *www.agile modeling.com*

Brandt, R.L. (1999). Porting the Web. *Upside,* September.

Cato, J. (2001). *User-centered Web design*. Harlow: Addison-Wesley.

Chan, H., Lee, R., Dillon, T., & Chang, E. (2001). *E-commerce: Fundamentals and applications*. Chichester, UK: John Wiley & Sons.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (1998). *Human-computer interaction* (2nd ed.). Harlow, UK: Prentice Hall.

Jennings, M. (2000). *Theory and models for creating engaging and immersive ecommerce Websites.* Paper presented at SIGPR 2000, Evanston, Illinois.

Kotze, P. (2000). Peopleware: Changing the mindset of computer science and engineering. *Inaugural lecture*, University of South Africa.

Kotze, P., & Johnson, C.W. (2001). *Human-computer interaction 1*. Study Guide for INF120-8, INF120-8/502/2001, University of South Africa.

Nielsen, J. (1996). *Top ten mistakes in Web design*. Retrieved June 25, 1999, from *www.useit.com/alerbox/9605.html*

Nielsen, J. (1999). *Usability as barrier to entry*. Retrieved December 8, 1999, from *www.useit.com/alertbox/991128.html*

Pressman, R.S. (2000). What a tangled Web we weave. *IEEE Software,* (January/February), 18-21.

Shneiderman, B. (1998). *Design the user interface: Strategies for effective human-computer interaction* (3rd ed.). Reading, MA: Addison-Wesley.

Turban, E., Rainer, R.K., & Potter, E.R. (2001). *Introduction to information technology*. New York: John Wiley & Sons.

KEY TERMS

Agile Development Methodologies: An approach used in building non-critical computer systems.

Electronic Catalogs: Vendors' catalogs offered either on CD-ROM or on the Internet to advertise and promote products and services.

Electronic Commerce: Uses some form of transmission medium through which exchange of information takes place in order to conduct business.

Interaction: How a user communicates, or interacts, with a computer. Interaction focuses on the flow of interaction, the dialog between person and computer, how input relates to output, stimulus-response compatibility and feedback mechanisms.

Interactive System: This is a system that supports communication in both directions, from user to computer and back again. A crucial property of any interactive system is *its support for human activity.*

Network: A telecommunication system that permits the sharing of resources such as computing power, software, input/output devices and data.

Transactional Software: Search engines for finding and comparing products; negotiating software; encryption and payment; ordering (front office) inventory and back office software.

Usability: ISO 9241-11 standard definition for usability identifies three aspects: (1) a specified set of users, (2) specified goals (tasks) that have to be measurable in terms of effectiveness, efficiency and satisfaction, and (3) the context in which the activity is carried out.

User Interface: Facilitates the communications between a user (a person) and an information system and may be tailored uniquely to an individual.

Educational Technology Standards

Michael O'Dea
University of Hull, UK

INTRODUCTION

The "holy grail" of e-learning is to enable individualized, flexible, adaptive learning environments that support different learning models or pedagogical approaches to learning to allow any Internet-connected user to undertake an educational program. It is also very highly desirable, from a more practical viewpoint, if this environment can also integrate into the wider MIS/student records system of the teaching institution.

A number of very different technologies in the past have been employed to try and achieve this aim, with varying degrees of success; see Hartley (1973), Muhlhausen (2003) and Okamoto and Hartley (2001) for good accounts of the development of ICT in education. However, one of the biggest stumbling blocks to date, hindering the widespread adoption of these technologies, has been the cost of developing these learning materials and their delivery systems, alongside an inability to reuse the materials.

Addressing these issues is now where much of the main research efforts within the e-learning field are focused, particularly in the developments of Learning Technology Standards.

The learning technology standardization process is leading the research effort in Web-based education. Standardization is needed for two main reasons: (1) educational resources are defined, structured and presented using various formats; (2) functional modules embedded in a particular learning system cannot be reused by another system in a straightforward way. (Anido-Rifon, Fernandez-Iglesias, Llamas-Nistal, Caeiro-Rodriguez and Santos-Gago, 2001)

Currently, a number of standards have been developed. For example, probably the three most commonly employed at present are IEEE's Learning Object Metadata—LOM (IEEE, 2001), ADL's Shareable Content Object Reference Model—SCORM (ADL, 2001) and the Open Knowledge Initiative – OKI (OKI, 2004). These standards, in turn, often incorporate other standards and specifications within them; for example, SCORM utilizes the IMS Content Packaging and Simple Sequencing specifications. The result of this is a plethora of acronyms and standards, which can prove confusing, even for some practitioners.

It is the aim of this article to clarify the aims, role and main functions of key current educational technology standards and to highlight the advantages they bring when learning environments are developed with them. The article will also address some of the aspects of e-learning not so well served by the standards and some of the current and future directions of research within the field.

The structure of the article is as follows: It will start with a brief background of e-learning, covering the main types of applications used to enable delivery of e-learning. The main section will be devoted to the considering the main learning technology standards, attempting in particular to highlight the many different standards and the roles they fulfill in enabling interoperability and compatibility between e-learning applications, but also to highlight the connections between the various standards. Finally, the article will examine some of the current issues of debate surrounding the standards.

E-LEARNING: A BRIEF BACKGROUND

E-learning is the use of the Web as a medium of delivery for educational ICT applications. The use of the Web potentially enables distance-independent, time-independent, computing platform-independent and classroom size-independent learning far more easily than alternative media of delivery, such as CD-ROM or broadcast multimedia.

In essence though, e-learning applications, like all educational ICT applications, strive to achieve two

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

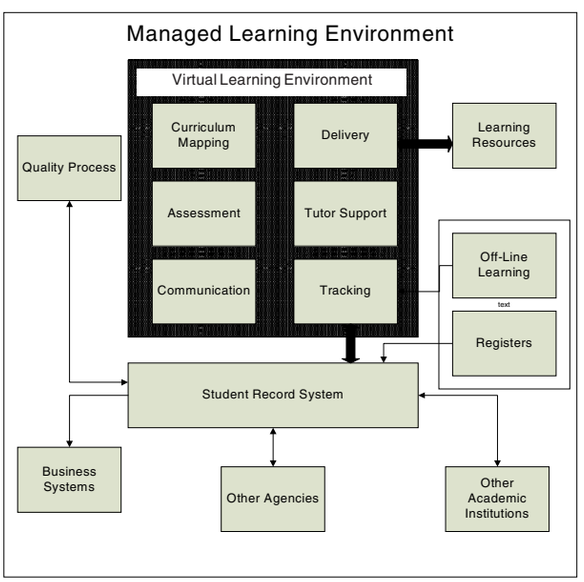
main aims: (1) present educational content, and (2) provide facilities and tools to enable learning.

The key technology of delivery of e-learning is the Learning Environment. Commercial examples of these include WebCT and Blackboard. Any brief perusal of e-learning-related literature will quickly reveal a number of terms used to describe Learning Environments. The most common of these are: Managed Learning Environment (MLE), Virtual Learning Environment (VLE), Learning Management System (LMS) and Learning Content Management System (LCMS). While it is technically correct to use any one of these terms to describe a learning environment, each has a subtle difference in meaning; therefore, it may be useful at this point to provide a brief definition.

MLEs and VLEs are terms used to describe the two main types of e-learning application.

MLEs can be considered to be enterprise level, large-scale e-learning applications. They aim to provide the whole range of information services an educational institution would require to enable and support the learning process and its operation (see Figure 1). Conole (2002) describes the main function of an MLE as to "integrate a VLE with a university's management systems" and goes on to note that this "might include a wide range of functional components … (such as) … administrative information about courses, resources, support and guidance, collaboration information, assessment and feedback, evaluation."

Figure 1. Structure of an MLE (adapted from JISC)



An MLE can, and normally does, include a VLE. A VLE deals with the actual delivery of the learning material or content, including assessment, tutor-to-learner communication and tracking of student progress and activity, as well as linking to any student record or Management Information System (which itself may or may not be part of an MLE). A VLE may also, often, include a content authoring facility. In essence, a VLE is the e-learning application that delivers the course to the learner. For those interested, Conole (2002) provides a good exposition of MLEs and VLEs in more detail.

In turn, a VLE may include the functions of either an LCMS or LMS or of both. There does appear to be some confusion in much of the literature in the use of the two terms. First, often they are used to describe the applications themselves, although it would appear that most definitions of them normally refer to functionality or the services that they provide. Second, the term LMS often is used as a blanket term to describe what others term an LCMS (see Jacobsen, 2002 for a good discussion of these issues). So to clarify this point, in this article, the following definitions will be used:

An LCMS manages the learning material and the learning process. Often they track individual learning progress. Typically, an LCMS will do the following: Course preparation, course delivery, tracking and itemizing of user details; for example, the number of times a user accesses a particular section of content and for how long.

An LMS manages the student and learning events that support the administration of the learning. The functionality described by an LMS may include: hosting the course catalog, administration of the course, such as scheduling of courses, tracking and reporting completions and results for individual students.

Jacobsen (2002) provides a much more detailed definition of the two terms, but has a very simple and effective description of the difference between an LMS and LCMS. An LMS "handles what takes place outside of the course" whilst an LCMS "handles what takes place within the (virtual) classroom."

THE LEARNING TECHNOLOGY STANDARDS

The role of educational technology standards in e-learning has been to attempt to enable interoperability and compatibility between VLEs, MLEs, LMSs and LCMSs. The standards attempt to provide interoperability specifications for certain key elements of the e-learning process and for the various support functions required to enable the process to take place. The fundamental concept, upon which virtually all current educational technology standards and specifications have been developed, is reusable chunks of information. These have variously been termed knowledge objects, content objects and, most commonly, learning objects. All of these concepts refer to small, self-contained objects of knowledge that offer the ability to enable reuse of content, enable modularized development of learning environments and applications, and enable standardized presentation of learning materials and content.

On top of this concept of there being a basic unit of learning material or content learning, technology standards have also focused on enabling semantic information to be attached to these, primarily to enable content management. The idea behind this is that the base learning material, the learning object, can be described in a richer way, can be interchanged more effectively and can be searched and stored more efficiently by a content management system. The work done on the standards that build on the basic learning object description can be seen as providing for the main semantic criteria of classification, metadata and ontologies.

A useful model to conceptualise semantic information is Berners-Lee's layered model of the semantic web, as shown in Figure 2. The WWW Consortium (W3C) describe the semantic web as "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in co-operation" (w3c.org). If we consider the objectives of learning standards, we can see a correspondence between the objectives of learning technology standards and the semantic web; that is, defining meaning to enable cooperation. We will see that the semantic definitions of learning material are the basic building blocks upon which most of the standards outlined later in this article rely.

Learning Objects

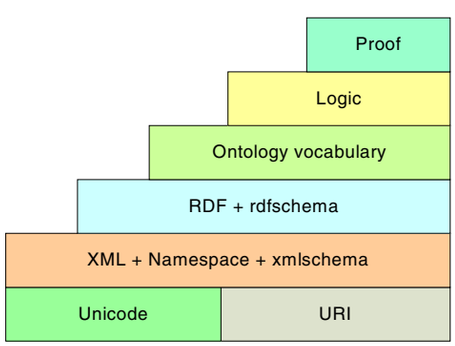
The base level standard for learning technology is the Learning Object (LO). The IEEE describes them as "Any entity, digital or non-digital, which can be used, re-used and referenced during technology-supported learning" (IEEE, 2001), while others have described them as chunks of learning content that can be combined to comprise teaching modules or courses. Each LO has the ability to communicate with a learning system that organises and manages it. This learning system may be a VLE or LCMS.

LOs have been designed not only to be reused, but also so that they can be easily delivered via variety of media, particularly the Web, and this enables any number of people to access and use them simultaneously. They provide a means for efficient development of computer-based, interactive, multimedia instruction. Examples of LOs suggested by the IEEE include: multimedia content, instructional content, instructional software, software tools, learning objectives, persons, organizations or events.

Learning Object Metadata

The first main standard to be developed upon the concept of LOs was the facility for each LO to have information—in particular, semantic information—attached to it that describes its contents. This information is called Learning Object Metadata (LOM). The aim of the LOM specification is to enable the reuse, search and retrieval of the LO's content and the integration of LOs with external systems.

Figure 2. The semantic web (Berners-Lee, w3.org)



A number of different LOM specifications exist. Each differs slightly, but significantly, in terms of the metadata it specifies and provides. The basic LOM specification is set out in the IEEE Learning Technology Standards Committee (LTSC) specification, LTSC 1484.12.1. This is based on the Dublin Core metadata schema and specifies a set of 47 metadata elements in nine categories (General, Lifecycle, Meta-metadata, Technical, Educational, Rights, Relation, Annotation and Classification) that have been selected to describe the most important aspects of a LO in order to enable reuse and interoperability.

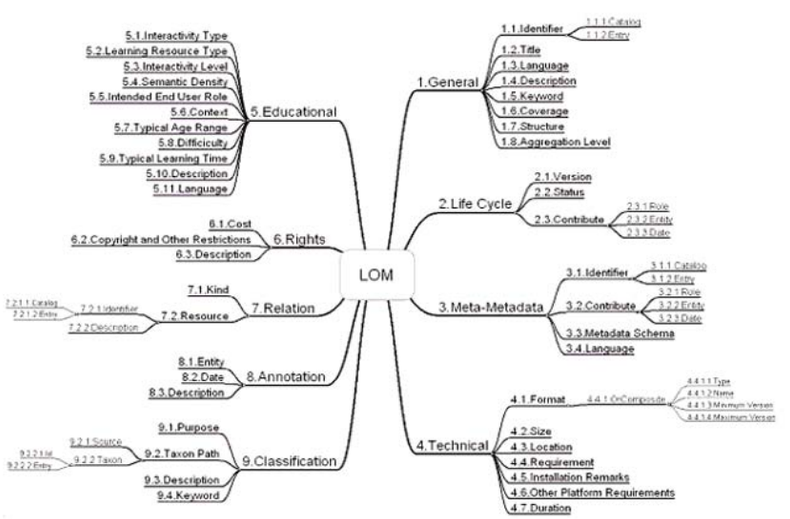
To date, the IEEE LOM standard has been specified in two formats: XML and RDF. To be of any use within a VLE or LCMS, representations or bindings compatible to the semantic web are required; that is, XML and RDF. These were specified by IEEE in 1484.12.3 and 1484.12.4, respectively, in 2002 and, at present, the XML representation has been ratified, but the RDF representation is still at the draft standard stage. Nilsson, Palmer and Brase (2003) provide a detailed exposition of this process and the standards' details.

However, there are a number of variants of the IEEE LOM specification, which have been based on the IEEE LOM standard but are essentially subsets of the IEEE standard with some extensions. Perhaps the most well established and widely used of these is the IMS Learning Resource Metadata specification (see IMS, 2004a), but other significant alternatives include the CanCore (CanCore, 2004) and SingCORE (E-Learning Competency Centre, 2004) Learning Resource metadata specification.

The rationale for the development of the IMS variant of LOM has been summarised by the MIT Libraries Metadata Advisory Group. They comment that the original IEEE LOM specification defined a very large set of items, and many organisations within the IMS community preferred to use a smaller subset of the elements. However, it was also felt that other elements were needed, so the IMS specification enables the extension of certain elements for "proprietary purposes" (MIT Libraries, 2004).

If we consider the current state of the standards' development in relation to Berners-Lee's layered model of the semantic web, we can see that these standards sit at layer 2 for the XML binding and layer 3 for the RDF binding; however, the LOM metadata element content itself is written in Unicode 3.0.1, also known as ISO/IEC 10646-1:2000, which sits at layer 1. In terms of functionality, the XML binding's main function is to enable reuse and interoperability of the LO. It provides the structural semantic information, while the RDF binding enables more effective search and retrieval of the LO by content management systems. For example, it provides the full semantic information needed for specialist ontologies to be developed to assist automated or semi-automated information retrieval to take place.

Figure 3. The elements and the structure of the IMS LOM standard (IMS)



Standards and Specifications that Pperate on Top of the Semantic Representations

Working alongside the main standards discussed above are a wide variety of standards and specifications that have also been developed to enable other aspects of reuse and interoperability between e-learning applications.

These can be considered to be of four types, dealing with:

1. The presentation of educational content
2. The provision of facilities and tools to enable learning
3. Enabling the management of learning programs
4. Combining other specifications or work alongside these to enable the development of learning environments

Standards or Specifications that Deal with the Presentation of Educational Content

While LOM enables reuse and integration of LOs, Content Packaging (CP) enables the aggregation of a number of LOs into a course or part of a course and allows this agglomeration of LOs to be treated as a unit. One of the shortcomings of the LO and LOM specifications is the lack of ability to enable different levels of granularity, or to hierarchically organise LOs, especially when they are required to be combined into modules or parts of modules. The IMS Content Packaging Specification partially addresses this issue.

A CP is described by the IMS as:

The objective of the IMS CP Information Model is to define a standardized set of structures that can be used to exchange content. These structures provide the basis for standardized data bindings that allow software developers and implementers to create instructional materials that interoperate across authoring tools, LMSs and run-time environments that have been developed independently by various software developers. (IMS, 2004b)

Closely related to this is SCORM's Content Aggregation Model. It goes a little further in its functionality by enabling the creation of Sharable Content Objects (SCOs), which comprise a set of LOs and which can be managed as a single entity by an e-learning learning environment, LMS or LCMS.

Standards or Specifications that Deal with the Provision of Facilities and Tools to Enable Learning

Content aggregation provides the platform for a VLE to deliver learning content. Operating alongside this within a VLE can be a number of other processes or functions, as shown in Figure 1. These aim to enable the learning process itself or provide tools to do so, in effect, to enable a degree of structure or process to the delivery of the modules within the e-learning application. Typically, this means managing the order of delivery, ensuring prerequisite modules have been taken before a student can undertake a module and enabling constraints on or specifying the modules that a student is able to attempt at any point in time during the course of study.

Both of the CP specifications discussed above provide no information on how the LOs they contain can be attempted or undertaken by the learner. They are pedagogy neutral; that is, their designs favour no particular pedagogical approach. In addition, they make no attempt to enable any pedagogical structure or strategies, but leave this to other parts of the learning environment. They do enable prerequisites to specified against LOs, but do nothing to implement them.

This role of structuring the delivery of the learning content and in enabling certain aspects of pedagogy to be implemented is currently handled by sequencing. The main sequencing specification currently in operation (mainly because it is included as part of the SCORM suite of specifications) is the IMS Simple Sequencing specification. This is described by the IMS as:

a method for representing the intended behaviour of an authored learning experience such that any learning technology system can sequence discrete learning activities in a consistent way. The specification defines the required behaviours and

functionality that conforming systems must implement. It incorporates rules that describe the branching or flow of instruction through content according to the outcomes of a learner's interactions with content. (IMS 2004c)

Essentially, Simple Sequencing defines how a learner can progress through the content of an e-learning set of activities. It does this by implementing an activity tree, which enables and constrains what options and in what order they can be attempted by the learner. So while Simple Sequencing is still pedagogy neutral, it is pedagogically aware.

Many commentators, such as Abdullah and Davis (2003), have argued that the Simple Sequencing specification is not flexible enough to fully implement an effective pedagogical model within an e-learning learning environment. One possible solution is the IMS Learning Design specification. This has been described "as one of the most significant recent developments in e-learning" (Dalziel, 2003). It goes a step further than the Simple Sequencing specification and enables the consideration of context on the learning process; in particular, it aims to enable the implementation of pedagogical approaches to the learning activity.

The IMS Learning Design specification supports the use of a wide range of pedagogies in online learning. Rather than attempting to capture the specifics of many pedagogies, it does this by providing a generic and flexible language. This language is designed to enable many different pedagogies to be expressed. The approach has the advantage over alternatives in that only one set of learning design and runtime tools then need to be implemented in order to support the desired wide range of pedagogies. (IMS 2004d)

The Learning Design specification provides facilities to assign people to roles, facilitate different types of interaction between the content and the learner, learners and other learners, and learners and tutors.

Standards or Specifications that Enable the Management of Learning Programs

Working alongside the specifications described above are other specifications that address many different aspects of e-learning learner support, in particular, the management and administration of students and courses. These include the IEEE LTSC Public and Private Information (PAPI) specification, which defines learner records and portfolios.

Another important specification aimed at administration is the IMS Enterprise Data Model. This provides standardized definitions for course structures and enables learning environments to schedule student activities from course structure definitions and the movement of courses between learning environments.

The AICC guidelines for interoperability of Computer Managed Instruction (CMI) systems propose a Web-based runtime environment scheme that enables learning content of different origins and formats to interface with a CMI compliant Web-based management system and launch on a browser and be controlled by the system. It also allows different learning resources to be managed by heterogeneous management systems. (IMS, 2004a)

Standards or Specifications that Combine Other Specifications or Work Alongside these to Enable the Development of Learning Environments

The standards and specifications discussed so far in this article can be considered to be micro-level specifications, based on LOs and the use and management of them. At the macro level, the specifications address the architecture of learning environments.

Probably the most significant of this type of standard is the Learning Technology Systems Architecture (LTSA). According to Conole, "the LTSA specification covers a wide range of systems (learning technology, computer-based training, electronic performance support systems, computer-assisted instruction, intelligent tutoring, education and training technology, metadata, etc.) and is intended to be pedagogically neutral, content-neutral, culturally neutral and platform-neutral" (Conole, 2003) It provides a "framework for ... promoting interoperability and portability by identifying critical system interfaces."

The LTSA is a five layer framework. It was developed as part of the IEEE P1484.1 project of the 1484.1 working group of IEEE LTSC. The main use

of the LTSA framework is as a model for framing interoperability issues.

However, in the field, the most significant of the macro-level standards is probably SCORM. Not only because alongside LOM it is the most well known and most widely employed and implemented of educational standards, but also because it comprises a collection of many of the standards discussed earlier in this article, plus a Run Time Environment that provides a foundation for learning environments to run these standards and so be developed upon a standard "engine." It currently provides the most comprehensive architecture upon which to develop an LCMS or LMS that is standards based.

SCORM has been developed by the Advanced Distributed Learning (ADL) and has been heavily supported and promoted by the United States government. Among others, it collects together a number of standards; notably, IMS Learning Resources Metadata, IMS Content Packaging, AICC CMI and IMS Simple Sequencing and combines them to provide a set of specifications that enable the "reuse of instructional components in multiple applications and environments regardless of the tools used to create them" (ADL, 2001). The SCORM architecture operates in a way that enables the separation of content from context-specific run-time constraints and the specification of common interfaces and data. By doing this it enables applications to provide different levels of functionality; for example, from simple LOM metadata editors and annotators to full-blown VLEs, and still be interoperable.

The second significant high-level set of specifications is the Open Knowledge Initiative (OKI). Developed at MIT and Stanford, this is an open-source reference system for Web-enabled education. Similar to SCORM, it provides a set of resources and an architecture designed to enable the development of easy-to-use, Web-based environments and for assembling, delivering and accessing educational resources. There does appear to be some overlap between SCORM and OKI; however, while the functionality of SCORM is very much focused on what can be considered the operation of a VLE, the OKI specification extends to address the functionality required for MLE development. Additionally, OKI is based on a set of APIs, the Open Source Interface Definition (OSID), which defines a set of programming interfaces, as opposed to SCORM, which is based on data definitions. This means in theory that they should work together side by side.

DISCUSSION AND CURRENT ISSUES

Understandably, SCORM and LOM receive most of the attention and scrutiny of the standards discussed, and they have provided the basis of a significant number of initiatives and projects worldwide, some with very substantial backing and budgets. However, while there can be no doubt that they have promoted and enabled interoperability, flexibility and reusability in the development of learning environments, there still are a number of issues of debate concerning them.

One area that has been the subject of much debate is the nature of the LO itself. This is significant, as the LO is very much the basis upon which virtually all of the learning technology standards and specifications discussed here have been built upon. Commentators such as Frierson (2003), Wiley (2001) and Rehak and Mason (2003) have highlighted that there still appears to be concern over just exactly what a LO is. "Different definitions abound, different uses are envisaged, and different sectors have particular reasons for pursuing their development. In this environment of uncertainty and disagreement, the various stakeholders are going off in all directions" (Rehak & Mason, 2003). Certainly this seems to be confirmed by the existence of competing LOM specifications; for example, IEEE, IMS, CanCore and SingCore, where the latter three have extended the base IEEE specification to include proprietary extensions to the LOM metadata and thus introduced a degree of incompatibility back into a specification that was supposed to enable interoperability and compatibility. This has proved in the past and is likely to continue to in the future, a significant barrier to interoperability and reuse of learning content.

Anido-Rifon (2001) notes also that the current LOM specifications' lack of internal descriptions of LOs is a problem. It can hinder interoperability in a number of ways; for example, if information needs to be adapted or utilised for other processes. This is significant if developers try to enable user-specifiable preferences as required for adaptive learning environments. We may also see, as a result of this, that further extensions to the LOM specifications may be developed or bridges to other metadata standards such as

MPEG7 are created, to enable the inclusion of the types of content that cannot be defined under the current LOM specifications into learning environments.

However, the most significant debate concerning learning technology standards and one which is applicable to all of the standards we have discussed so far, is what Wiley (2001) describes as the "decontextualisation" of learning that results as a consequence of using the learning technology standards approach to developing educational tools. Commentators such as Wiley (2001), Freisen (2003) and Rehak (2003) argue that the pedagogical-neutral model upon which all of the technologies are based removes all aspects of instructional context from the learning process, and that this is something that is at odds with modern educational theory. Wiley notes, "while economically sensible, the drive towards decontextualisation may actually be counter productive from the standpoint of student learning" (Wiley, 2001). A sentiment backed up by Rehak, who, commenting on SCORM in particular, states that it is based on "a limited pedagogical model unsuitable for some environments" (Rehak, 2003).

One standard that is the subject of particular criticism is Simple Sequencing. Simple Sequencing is based on what is described as "a single user interaction model of behaviour" (Rehak, 2003), which according to Rehak, "does not easily accommodate multiple user environments, especially those requiring different courses of action for different users" (Rehak, 2003). The IMS Learner Design specification is intended to be an improvement on Simple Sequencing and was specifically designed to enable the ability to implement some aspects of pedagogy into learning environments. However, although Learning Design is a significant advance in functionality over Simple Sequencing, there still are concerns over the Learning Design concept and its implementation. Some question the actual reusability of Learning Design specifications. Downes (2003) goes as far as to comment that, "Learning design and reusability are incompatible." Basically, his argument is that learning design specifications need to be so tailored to individual environments or contexts that, as a result, they are unable to be of general applicability.

So what hope, then, is there for educational technology standards? The future may lie in the semantic web. It was noted earlier how LOM and its XML and RDF representations sit at layers 2 and 3 of Berners-Lee's semantic web stack (see Figure 2). Many of the standards outlined above aim to provide semantic information to learning content, interpret the semantic information of the learning content to enable the e-learning process or provide structure to the semantic information, yet none of the standards could be considered to sit at any of the layers above 3. This is possibly because almost all are based on the XML representation of LOM. However, as yet, little work has been done on educational ontologies or vocabularies, the type of specifications that would sit at layer 4 of the semantic web stack and would be based on the RDF representations of LOM and so interpret meaning of content. Once we have these in place, it gives the ability to develop domain models, user models and possibly most significantly pedagogical models, giving the option of providing context and the ability to implement pedagogy to learning environments based on standards (see O'Dea, Huang & Mille, 2003) and even fully adaptive learning environments.

CONCLUSION

This article has shown that during the past 15 or so years there have been significant developments in the field of education technology standards. There is now a comprehensive and well-specified suite of standards and specifications that address many of the aspects of learning environment functionality, both for VLEs and, to a lesser extent, MLEs. It is fair to say that the aims of interoperability, reuse and flexibility that initiated the efforts in the field are much further on the way to being achieved.

However, the debate regarding the pedagogically neutral model upon which the standards are based is a significant one. Wiley (2001) places it in perspective:

Many of the problems ... only actually become problems as desired learning outcomes climb further up Bloom's taxonomy. Issues of decontextualisation, mediation and socialization are all but non-issues when the desired learning outcome is acquisition ... of information and the assessment is recall. However, to the degree to which higher-order learning outcomes

(such as synthesis and evaluation) are called for, or to which an explicit emphasis would be placed on transfer from an institutional context into a later performance context, we believe these issues become critical problems. (Wiley, 2001)

A potential solution to these issues exists in the semantic web, in using ontologies as the basis for the development of domain models, user models and pedagogical models, which in turn can provide the ability to implement context into learning based on educational technology standards.

If we do not address these issues, we run the risk of the issues that Downes (2003) identifies: Instead of reusable content, we may have to move to models based on disposable content, which then goes against the one of the major foundations of the whole move towards standards and standardization.

In short, yes, the move towards standards for learning technology has resulted in progress towards the aims for which it was commenced—reusability, interoperability and heterogeneity. However, we are not there yet, and there still is a long way to go in certain areas.

REFERENCES

Abdullah, N., & Davis, H. (2003, August 26-30) *Is simple sequencing simple adaptive hypermedia?* ACM Sigweb 14th Conference on Hypertext and Hypermedia, Nottingham, UK.

ADL. (2001). The SCORM Overview ADL. Retrieved 26/3/04 from *www.adlnet.org/index.cfm*

Anido-Rifon, L., Fernandez-Iglesias, M.J., Llamas-Nistal, M., Caeiro-Rodriguez, M., & Santos-Gago, J. (2001). A component model for standardized Web-based education. *ACM Journal of Educational Resources in Computing, 1*(2).

CanCore. (2004). Canadian core learning resource metadata application profile. Retrieved March 26, 2004, from *www.cancore.ca/indexen.html*

Conole, G. (2002). Systematising learning and research information. *Journal of Interactive Media in Education,* (7).

Dalziel, J. (2003). Implementing learning design: The Learning Activity Management System (LAMS). *ASCLITE 2003.*

Downes, S. (2003). Design standards and reusability. Retrieved March 26, 2004, from *www.downes.ca/cgi-bin/website/*

E-Learning Competency Centre. (2004). SingCORE. Retrieved March 26, 2004, from *www.ecc.org.sg*

Fisher, M., & Sheth, A. (2003). Semantic enterprise content management. In M. Singh (Ed.), *Practical handbook of Internet computing*. Boca Raton: CRC Press.

Friesen, N. (2004). Three objections to learning objects. In R. McGreal (Ed.), *Online education using learning objects.* London: Routledge/Falmer.

Gibbons, A.S., & Fairweather, P.G. (2000). Computer based instruction. In S. Tobias & J.D. Fletcher, (Eds.), *Training and retraining: A handbook for business, industry, government and the military*. New York: MacMillan.

Hartley, R. (1973). The design and evaluation of an adaptive teaching system. *Internet Journal of Man-Machine Studies, 5*, 421-436.

IEEE. (2001). IEEE Learning Technology Standards Committee IEEE P1484.12 Learning Object Metadata Working Group; WG12. Retrieved March 26, 2004, from *www.imsglobal.org/metadata/index.cfm*

IMS. (2004a). IMS Learning Resource Metadata Specification. Retrieved March 26, 2004, from *www.imsglobal.org/metadata/index.cfm*

IMS. (2004b). IMS Content Packaging Information Model - Version 1.1.3 Final Specification. Retrieved March 26, 2004, from *www.imsglobal.org/content/packaging/*

IMS. (2004c). IMS Simple Sequencing Specification. Retrieved March 26, 2004, from *www.imsglobal.org/simplesequencing/index.cfm*

IMS. (2004d). IMS Learning Design Specification. Retrieved March 26, 2004, from *www.imsglobal.org/learningdesign/index.cfm*

Ismail, J. (2002). The design of an e-learning system: Beyond the hype. *The Internet and Higher Education*, *4*, 329-336.

Jacobsen, P. (2002). LMS vs LCMS. *E-Learning,* June.

MIT Libraries. (2004). IMS Learning Resource Metadata Information Model. Retrieved March 26, 2004, from *http://libraries.mit.edu/guides/subjects/metadata/standards/ims.html*

Muhlhauser, M. (2003, December 10-12). Multimedia software for e-learning: An old topic seen in new light. *Proceedings of IEEE 5th International Symposium on Multimedia Software,* Taichung, Taiwan.

Nilsson, M., Palmer, M., & Brase, J. (2003). *The LOM RDF binding – Principles and implementation.* The 3rd Annual Ariadne Conference, Katholieke Universiteit Leuven, Belgium.

O'Dea, M., Huang, W., & Mille, A. (2003, December 10-12). ConkMeL: A Contextual Knowledge Management framework to support intelligent multimedia e-learning. *Proceedings of IEEE 5th International Symposium on Multimedia Software Engineering,* Taichung, Taiwan.

Okamoto, T., & Hartley, R. (2002). Innovations in learning technology. *Educational Technology and Society,* (4), 8-10.

OKI. (2004). The Open Knowledge Initiative. Retrieved March 26, 2004, from *http://web.mit.edu/oki*

Rehak, D., & Mason, R. (2003). Keeping the Learning in Learning Objects. In A. Littlejohn (Ed.), *Reusing online resources: A sustainable approach to eLearning*. London: Kogan Page.

Wiley, D. (2003). Learning objects: Difficulties and opportunities. Retrieved March 26, 2004, from *http://wiley.ed.usu.edu/docs/lo_do.pdf*

KEY TERMS

Learning Object Metadata (LOM): LOM is semantic information attached to Learning Objects. There are a number of LOM standards. The main aim of the LOM specification is to enable the reuse, search and retrieval of the LOs' content and the integration of LOs with external systems.

Learning Objects (LO): Chunks of learning content that can be combined to comprise teaching modules or courses.

Managed Learning Environments (MLE): Managed Learning Environments can be considered enterprise-level, large-scale e-learning applications. They aim to provide the whole range of information services an educational institution would require to enable and support the learning process and its operation. A key part of the functionality provided by an MLE is connectivity to all elements of an educational institution's information systems.

Open Knowledge Initiative (OKI): OKI is an open-source reference system for Web-enabled education. It provides a set of resources and an architecture designed to enable the development of easy-to-use Web-based environments and for assembling, delivering and accessing educational resources.

Semantic Web: "An extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in co-operation" (w3c.org). It is based on Metadata information that is attached to documents. This metadata is designed to be machine readable and is expressed in XML and RDF format.

Shareable Content Object Reference Model (SCORM): SCORM collects a number of standards, notably, IMS Learning Resources Metadata, IMS Content Packaging, AICC CMI and IMS Simple Sequencing and combines them to provide a set of specifications that enable the "reuse of instructional components in multiple applications and environments regardless of the tools used to create them" (ADL 2001).

Virtual Learning Environments (VLE): A VLE deals with the actual delivery of the learning material or content, including assessment, tutor-to-learner communication and tracking of student progress and activity, as well as linking to any student record or Management Information System. A VLE may also, often, include a content authoring facility. In essence, a VLE is the e-learning application that delivers the course to the learner.

Efficient Method for Image Indexing in Medical Application

Richard Chbeir
University of Bourgogne, France

INTRODUCTION

In last two decades, image retrieval has seen a growth of interests in several domains. As a result, a lot of work has been done in order to integrate it in the standard data processing environments (Rui, Huang, & Chang, 1999; Smeulders, Gevers, & Kersten, 1998; Yoshitaka & Ichikawa, 1999). To retrieve images, different methods have been proposed in the literature (Chang & Jungert, 1997; Guttman, 1984; Lin, Jagadish, & Faloutsos, 1994). These methods can be grouped into two major approaches: metadata-based and content-based approaches. The metadata-based approach uses alphanumeric attributes and traditional techniques to describe the context and/or the content of the image such as title, author name, date, and so on. The content-based approach uses image processing algorithms to extract low-level features of images such as colors, textures, and shapes. Image retrieval using these features is done by methods of similarity and hence is a non-exact matching.

The requirement of each method depends on the application domain. In this paper, we address the domain of medicine where image retrieval in particular is very complex and should consider:

- Both content-based and metadata representations of images and salient objects. This guarantees a pertinent integration of all the aspects of image in order to capture pertinent information and to assure the relevance of all query types (Chbeir, Atnafu, & Brunie, 2002).
- High-precision description of images. For example, the spatial data in surgical or radiation therapy of brain tumors is decisive because the location of a tumor has profound implications on a therapeutic decision (Chbeir, Amghar, & Flory, 2001; Chbeir et al., 2002). Furthermore, it is crucial to distinguish between similar situations. Figure 1 shows two different images of three salient objects that are traditionally described by the same spatial relations in both cases: topological relations: a1 Touch a2, a1 Touch a3, a2 Touch a3; and directional relations: a1 Above a3, a2 Above a3, a1 Left a2.
- The evolutionary aspect of image content (Chbeir, Amghar, Flory, & Brunie, 2001) such as tumor development in brain (Figure 2), virus changes, and so on. The detection of the evolutionary aspects of objects (displacement, deformation, contraction, rotation, etc.) can significantly help physicians to establish an appropriate diagnosis or to make a therapeutic or surgical decision. An example for such a query is: "Find treatments of lesion detected inside brain images where a size increasing has been observed at every examination between time t and t+n".

In this article, we address the spatial and evolutionary issues of images. We propose a novel method that considers different types of relations. This method allows providing a highly expressive and powerful mechanism for indexing images.

The rest of this article is organized as follows: the next section is devoted to detail the related work. In the following section, we define our method of

Figure 1. Two different spatial situations

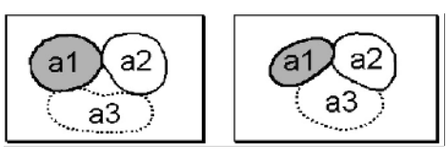
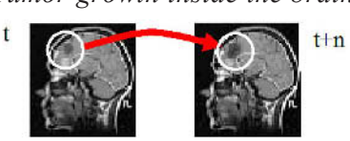


Figure 2. Tumor growth inside the brain



Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

computing the different relations and we show how image indexing can be done. The subsequent section demonstrates how our method can adequately index medical images. Finally, we conclude and give future work orientations.

RELATED WORK

The problem of image retrieval is strongly related to image representation. Computing relations between either salient objects, shapes, points of interests, etc. have been widely used in image representation such as R-tree and its variants (Beckmann, 1990; Guttman, 1984), hB-tree (Lomet & Salzberg, 1990), ss-tree (White & Jain, 1996), TV-tree (Lin et al., 1994), 2D-String and its variants (Chang & Jungert, 1997; Chang & Jungert, 1991; Chang, Shi, & Yan, 1987), and so on. Spatial relations are mostly used for indexing and retrieval purposes for its automatic detection capability.

Three major types of spatial relations are generally proposed in image representation (Egenhofer, Frank, & Jackson, 1989):

- Metric relations measure the distance between salient objects (Peuquet, 1986). For instance, the metric relation "far" between two objects A and B indicates that each pair of points A_i and B_j has a distance grater than a certain value d.
- Directional relations describe the order between two salient objects according to a direction, or the localisation of salient object inside images (El-kwae & Kabuka, 1999). In the literature, fourteen directional relations are considered:
  - **Strict:** north, south, east, and west.
  - **Mixture:** north-east, north-west, south-east, and south-west.
  - **Positional:** left, right, up, down, front and behind.

Directional relations are rotation variant and there is a need to have referential base. Furthermore, directional relations do not exist in certain configurations.

- Topological relations describe the intersection and the incidence between objects. Egenhofer and Herring (1991) have identified six basic relations: disjoint, meet, overlap, cover, contain, and equal. Topological relations present several characteristics that are exclusive to two objects (i.e., there is one and only one topological relation between two objects). Furthermore, topological relations have absolute value because of their constant existence between objects. Another interesting characteristic of topological relations is that they are transformation, translation, and scaling invariant.

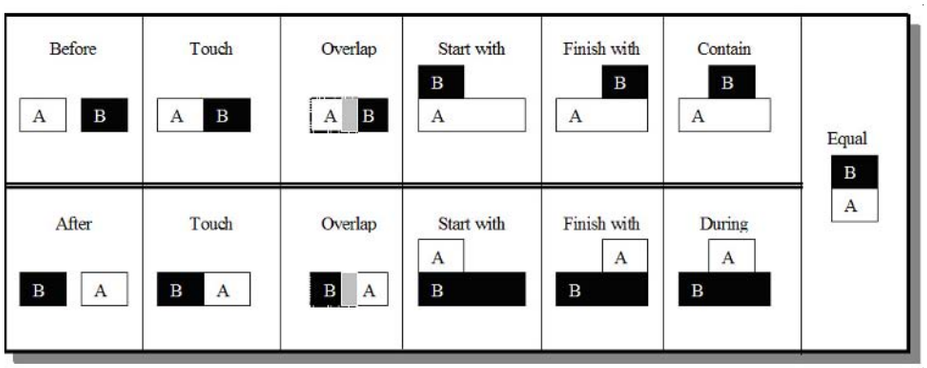
In spite of all proposed work to represent complex visual situations, several shortcomings exist in the methods of spatial relation computations. Particularly, traditional methods do not have the required expressive power to distinguish between similar situations in critical domains such as in medicine.

On the other hand, the evolution of image content needs to be taken into consideration in several domains. Any evolution needs to consider time and thus temporal relations. For that reason, two paradigms are proposed in the literature in order to compute temporal relations:

- The first paradigm consists of representing the time as a set of instants: $t_1,..t_i,..t_n$. Traditionally, only three temporal relations are possible between two objects: before, its symmetric relation after, and equal.
- The second paradigm considers the time as a set of intervals $[t_i, t_j]$. Allen relations (Allen, 1983) are often used to represent temporal relations between intervals. Allen proposes 13 temporal relations (Figure 3), in which six are symmetrical.

For instance, in geographic applications, spatio-temporal queries (Bonhomme, Trepied, Aufaure, & Laurini, 1999) are used more and more to study the translation of a mobile object, the evolution of spatial objects, and so on. In the medical domain, images are evolutionary in nature. Their content evolution description can provide an important support for the treatment of diseases. To the best of our knowledge, the evolutionary content of medical images was only studied by Cárdenas, Ieong, Taira, Barker, and Breant, (1993); Chu, Hsu, Cárdenas et al. (1998). In Cardenas et al. (1993), the authors study the development of

Figure 3. Allen relations



bones structures. They define a temporal evolutionary data model (TEDM). It extends traditional constructs. However, several evolutionary situations (such as deformation, expansion, contraction, etc.) are not considered.

In this paper, we present our indexing method that is capable to represent spatial relations in highly expressive manner and to provide efficient technique to detect evolutionary content of images.

META-MODEL OF RELATIONS

Our proposal represents a generalized extension of the 9-Intersection model and its variants (Egenhofer & Franzosa, 1991). It provides a method for computing not only topological relations but also other types of relations with higher precision.

The idea is to identify relations of two values (or instances) on the basis of the same feature (such as shape, position, time, etc.). For example, the shape feature expresses spatial relations, the time feature provides temporal relations, and so on. To identify a relation between two feature values, we use an intersection matrix between several sets defined below.

Definition of Intersection Sets

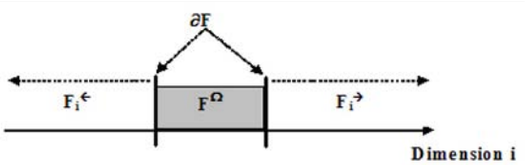
Let us first consider a feature F. We define its intersection sets as follows:

- The interior F^{Ω} contains all elements that cover the interior or the core of F. In particular, it contains the barycentre of F. The definition of this set has a great impact on the other sets. F^{Ω} may be empty ($\varnothing$).
- The boundary ∂F contains all elements that allow determining the frontier of F. ∂F is never empty ($\neg\varnothing$).
- The exterior F^{-} is the complement of $F^{\Omega} \cup \partial F$. It contains at least two elements: the minimum value and the maximum value. F^{-} can be divided into several disjoint exterior subsets. This decomposition depends on the number of the feature dimensions.

If the feature has only one dimension i (such as the acquisition time of a frame in a video), two exterior intersection subsets are defined (Figure 4):

- $F_i^{\leftarrow}$ (or inferior) contains elements of F^{-} that do not belong to any other intersection set and inferior to ∂F elements on the basis of i dimension.
- $F_i^{\rightarrow}$ (or superior) contains elements of F^{-} that do not belong to any other intersection set and superior to ∂F elements on the basis of i dimension.

Figure 4. Intersection sets of one-dimensional feature



If we consider a feature of two dimensions i and j (as the 2D shape of a salient object in Figure 5), we then define four exterior intersection subsets:

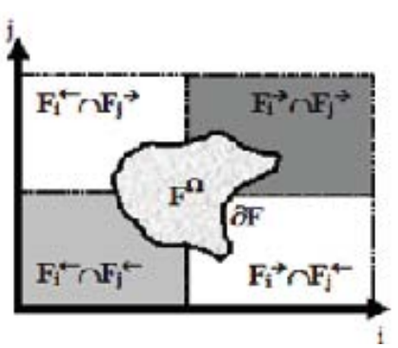
- $F_i^{\leftarrow} \cap F_j^{\leftarrow}$ (or inferior) contains elements of F^- that do not belong to any other intersection set and inferior to F^{Ω} and ∂F elements on the basis of i and j dimensions.
- $F_i^{\rightarrow} \cap F_j^{\rightarrow}$ (or superior) contains elements of F^- that do not belong to any other intersection set and superior to F^{Ω} and ∂F elements on the basis of i and j dimensions.
- $F_i^{\leftarrow} \cap F_j^{\rightarrow}$ contains elements of F^- that do not belong to any other intersection set and inferior to F^{Ω} and ∂F elements on the basis of i dimension, and superior to F^{Ω} and ∂F elements on the basis of j dimension.
- $F_i^{\rightarrow} \cap F_j^{\leftarrow}$ contains elements of F^- that do not belong to any other intersection set and superior to F^{Ω} and ∂F elements on the basis of i dimension, and inferior to F^{Ω} and ∂F elements on the basis of j dimension.

More generally and using the same reasoning, we are able to determine intersection sets (2n) of n dimensional feature.

In addition, we use a tolerance degree in the feature intersection sets definition in order to represent separations between sets and to provide a simple flexibility parameter for the computing process. For this purpose, we use two tolerance thresholds:

- Internal threshold ε^i: defines the distance between F^{Ω} and ∂F,
- External threshold ε^e: defines the distance between subsets of F^-.

Figure 5. Intersection sets of polygonal shape



Relations Computing Via Intersection Matrix

To calculate relation between two feature values, we establish an intersection matrix of their corresponding feature intersection sets. Matrix cells have binary values:

- 0 whenever intersection between sets is empty
- 1 otherwise

For two-dimensional feature (such as the shape) of two values A and B, we obtain the following intersection matrix.

The indexing relation between two values is expressed then by a binary value which is the juxtaposition of each raw of the corresponding intersection matrix. This is very important for indexing purposes.

Indexing Images

To index an image, we proceed as follows. First, we identify the spatial relations between the whole image and its salient objects. This allows determining the relative position of each object inside the image. In this case, the shape is used as a feature. The image is considered as a special object[1] that encloses all salient objects. We compute the intersection matrix of each pair of Image/Salient object. Figure 7 shows the result of this step on a 2D image I: three different spatial relations between the image I and its salient objects (a1, a2, and a3) are identified. It is important to mention that these relations are traditionally described by only one expression (i.e. northwest). We can see that, even at this level, our method provides a higher expression capability.

The second step consists of computing spatial relations between each pair of salient objects. Our method allows combining both directional and topological relation into one binary relation. The directional and topological relations are replaced by an expressive binary spatial relation between two salient objects. Figure 8 shows the result of this step where a distinguished[2] spatial relation is identified for each pair of salient objects of the image I.

Figure 6. Intersection matrix of two values A and B on the basis of two-dimensional feature

$A^{\Omega}\cap B^{\Omega}$	$A^{\Omega}\cap\partial B$	$A^{\Omega}\cap B_1^{+}\cap B_2^{+}$	$A^{\Omega}\cap B_1^{+}\cap B_2^{\rightarrow}$	$A^{\Omega}\cap B_1^{\rightarrow}\cap B_2^{\rightarrow}$	$A^{\Omega}\cap B_1^{\rightarrow}\cap B_2^{+}$
$\partial A\cap B^{\Omega}$	$\partial A\cap\partial B$	$\partial A\cap B_1^{+}\cap B_2^{+}$	$\partial A\cap B_1^{+}\cap B_2^{\rightarrow}$	$\partial A\cap B_1^{\rightarrow}\cap B_2^{\rightarrow}$	$\partial A\cap B_1^{\rightarrow}\cap B_2^{+}$
$A_1^{+}\cap A_2^{+}\cap B^{\Omega}$	$A_1^{+}\cap A_2^{+}\cap\partial B$	$A_1^{+}\cap A_2^{+}\cap B_1^{+}\cap B_2^{+}$	$A_1^{+}\cap A_2^{+}\cap B_1^{+}\cap B_2^{\rightarrow}$	$A_1^{+}\cap A_2^{+}\cap B_1^{\rightarrow}\cap B_2^{\rightarrow}$	$A_1^{+}\cap A_2^{+}\cap B_1^{\rightarrow}\cap B_2^{+}$
$A_1^{+}\cap A_2^{\rightarrow}\cap B^{\Omega}$	$A_1^{+}\cap A_2^{\rightarrow}\cap\partial B$	$A_1^{+}\cap A_2^{\rightarrow}\cap B_1^{+}\cap B_2^{+}$	$A_1^{+}\cap A_2^{\rightarrow}\cap B_1^{+}\cap B_2^{\rightarrow}$	$A_1^{+}\cap A_2^{\rightarrow}\cap B_1^{\rightarrow}\cap B_2^{\rightarrow}$	$A_1^{+}\cap A_2^{\rightarrow}\cap B_1^{\rightarrow}\cap B_2^{+}$
$A_1^{\rightarrow}\cap A_2^{\rightarrow}\cap B^{\Omega}$	$A_1^{\rightarrow}\cap A_2^{\rightarrow}\cap\partial B$	$A_1^{\rightarrow}\cap A_2^{\rightarrow}\cap B_1^{+}\cap B_2^{+}$	$A_1^{\rightarrow}\cap A_2^{\rightarrow}\cap B_1^{+}\cap B_2^{\rightarrow}$	$A_1^{\rightarrow}\cap A_2^{\rightarrow}\cap B_1^{\rightarrow}\cap B_2^{\rightarrow}$	$A_1^{\rightarrow}\cap A_2^{\rightarrow}\cap B_1^{\rightarrow}\cap B_2^{+}$
$A_1^{\rightarrow}\cap A_2^{+}\cap B^{\Omega}$	$A_1^{\rightarrow}\cap A_2^{+}\cap\partial B$	$A_1^{\rightarrow}\cap A_2^{+}\cap B_1^{+}\cap B_2^{+}$	$A_1^{\rightarrow}\cap A_2^{+}\cap B_1^{+}\cap B_2^{\rightarrow}$	$A_1^{\rightarrow}\cap A_2^{+}\cap B_1^{\rightarrow}\cap B_2^{\rightarrow}$	$A_1^{\rightarrow}\cap A_2^{+}\cap B_1^{\rightarrow}\cap B_2^{+}$

Figure 7. The spatial relations between the image and its 3 salient objects

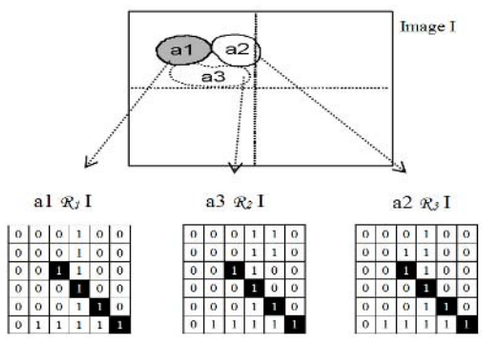
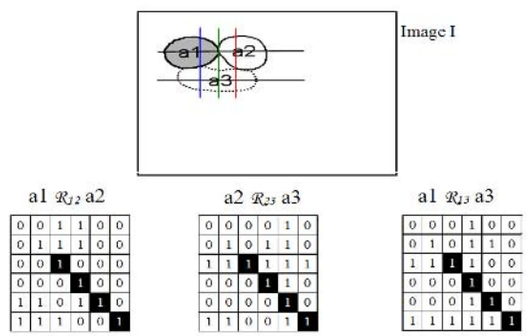


Figure 8. The spatial relations between salient objects



As a result, the image I is indexed as:

$$(a1\ R_1\ I + a2\ R_2\ I + a3\ R_3\ I) + (a1\ R_{12}\ a2 + a2\ R_{23}\ a3 + a1\ R_{13}\ a3)$$

Indexing Evolutionary Content

To index evolutionary content of images, we compare some of their features values. We consider that calibrating techniques are already applied before comparing. In Chbeir et al. (2001b), we identified several types of evolutionary possibilities in the medical domain. In this paper, we address only two major possibilities: the displacement and the transformation of salient object.

The displacement represents the evolution of salient object position between two images. For example, the displacement of a fibrinocruorique clot through a deep vein of the lower limb is very common in medicine. The displacement of a salient object can be captured by comparing its relative spatial relations with images, and then its spatial relations with other salient objects. Using our method, even similar and complex spatial situations are detected (see following below).

The transformation represents the evolution of an object shape between two images. For example, the expansion of tumor size inside the left lobe of a brain is an example of this type of evolution. The transformation can be expressed as deformation, contraction, stability, and so on. To detect salient object transformation, we consider the surface as a feature. In this case, the exterior set F^{-} is considered as indivisible. We determine then the relation between the surface changes of the same salient object detected in two images. In the medical domain, it is obvious that the two images must belong to the same patient. Figure 9 snapshots some of the possible evolutionary relations that can be detected.

APPLICATION EXAMPLE

Let us consider a medical image with three salient objects a1, a2, and a3 (Figure 10) where shapes have changed during a period of time. These two situations resemble similar but actually different. The distinction between them must be well expressed especially in the application domains that require precision (such as in medicine). As shown in Figure 10, our method

Figure 9. Several evolutionary relations detected

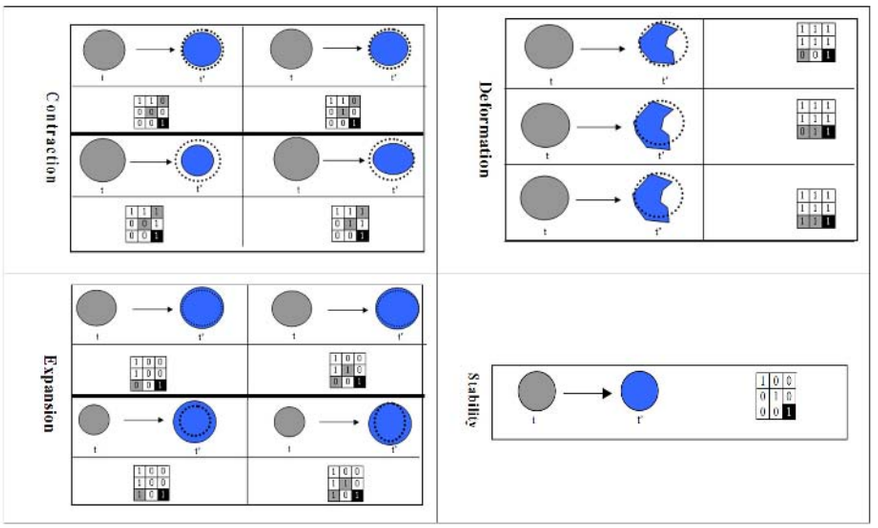
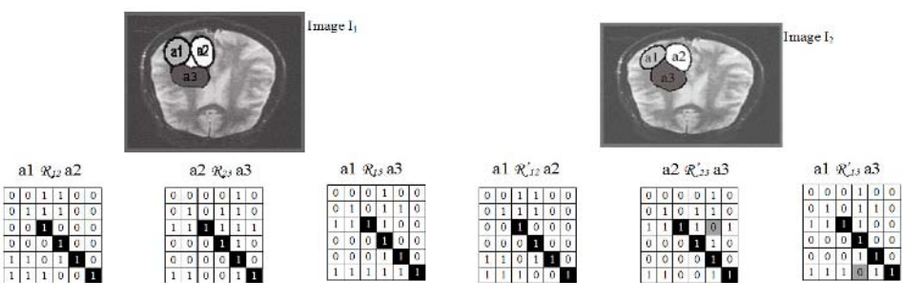


Figure 10. Using our method, the identification of similar spatial situations is possible



is capable to distinguish the different situations that are traditionally impossible to detect. The relations R_{12} and R'_{12} are equal but both relations R_{23} / R'_{23}, and R_{13} / R'_{13} are clearly distinguished.

Furthermore, our indexing method allows detecting the salient objects transformations: the deformation of a3, and the contraction of both a1 and a2.

CONCLUSION

We presented the domain of spatio-temporal image indexing and an original method capable of considering different types of relations in image representation. With our method, it is possible to homogenize, reduce and optimize the relations in image description models (Chbeir et al., 2002). In this article, we showed how to provide a highly expressive power to spatial relations that can be applied to describe images and then to formulate complex visual queries. We also showed how to detect evolutionary content of images. The example in medical domain demonstrates how image indexing can be improved.

Currently, we are working on integrating such indexing method in our prototype EMIMS (Chbeir et al., 2001a; Chbeir et al., 2002). Future work includes considering indexing of low-level features (color, texture, etc.) and more intense experiments in complex environment where large number of feature dimensions (2,5D[3] and 3D images) and salient objects exist. Our method will also be studied to see if it can used in datagrid computing.

REFERENCES

Allen, J.F. (1983). Maintaining knowledge about temporal Intervals. *Communications of the ACM, 26*(11), 832-843.

Beckmann, N. (1990). The R*-tree: An efficient and robust access method for points and rectangles. *SIGMOD Record, 19*(2), 322-331.

Bonhomme, C., Trepied, C., Aufaure, M.A., & Laurini, R. (1999). A visual language for querying spatio-temporal database. *Proceedings of the 7th International Symposium on GIS, ACMGIS' 99*, Kansas City, (pp. 34-39).

Cardenas, A.F., Ieong, I.T., Taira, R.K., Barker, R., & Breant, C. (1993). The knowledge-based object-oriented PICQUERY+ language. *IEEE-Transactions-on-Knowledge-and-Data-Engineering, 5*(4), 644-657.

Chang, S.K. & Jungert, E. (1991). Pictorial data management based upon the theory of symbolic projections. *Journal of Visual Languages and Computing, 2*(3), 195-215.

Chang, S.K. & Jungert, E. (1997). Human- and system-directed fusion of multimedia and multimodal information using the sigma-tree data model. *Proceedings of the 2nd International Conference on Visual Information Systems*, San Diego, (pp. 21-28).

Chang, S.K., Shi, Q.Y., & Yan, C.W. (1987). Iconic indexing by 2-D strings. *IEEE-Transactions-on-Pattern-Analysis-and-Machine-Intelligence, PAMI-9*(3), 413-428.

Chbeir, R., Amghar, Y., & Flory, A. (2001a). A prototype for medical image retrieval. *International Journal of Methods of Information in Medicine, Schattauer*, 3, 178-184.

Chbeir, R., Amghar, Y., Flory, A., & Brunie L. (2001b). A hyper-spaced data model for content and semantic-based. *Proceedings of ACS/IEEE International Conference on Computer Systems and Applications*, Beirut, Lebanon (pp. 161-167).

Chbeir, R., Atnafu, S., & Brunie, L. (2002). Image data model for an efficient multimcriteria query: A case in medical databases. *The 14th International Conference on Scientific and Statistical Database Management, IEEE Computer Society Press*, Edinburgh, Scotland, July 24-26 (pp. 165-174).

Chu, W.W., Hsu, C.C., Cárdenas, A.F., et al. (1998). Knowledge-based image retrieval with spatial and temporal constructs. *IEEE Transactions on Knowledge and Data Engineering, 10*(6), 872-888.

Egenhofer, M. & Franzosa, R. (1991). Point-set topological spatial relations. *International Journal of Geographical Information Systems*, *5*(2), 161-174.

Egenhofer, M. & Herring, J. (1991). *Categorising binary topological relationships between regions, lines, and points in geographic databases, a framework for the definition of topological relationships and an algebraic approach to spatial reasoning within this framework.* Technical Report 91-7, National center for Geographic Information and Analysis, University of Maine, Orono.

Egenhofer, M., Frank, A. & Jackson, J. (1989). A topological data model for spatial databases. Symposium on the Design and Implementation of Large Spatial Databases, Santa Barbara, CA. *Lecture Notes in Computer Science*, *409*, 271-286.

El-kwae, M.A. & Kabuka, M.R. (1999). A robust framework for content-based retrieval by spatial similarity in image databases. *ACM Transactions on Information Systems, 17*(2), 174-198.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. *SIGMOD Record, 14*(2), 47-57.

Lin, K.I., Jagadish, H.V., & Faloutsos, C. (1994). The TV-tree: An index structure for high dimensional data. *Journal of Very Large DataBase, 3*(4), 517-549.

Lomet, D.B. & Salzberg, B. (1990). The hb-tree: A multiattribute indexing method with good guaranteed performance. *ACM Transactions on Database Systems, 15*(4), 625-658.

Peuquet, D. J. (1986). The use of spatail relationsships to aid spatial database retrieval. *Proceedings of the 2nd International Symposium on Spatial Data Handling*, Seattle (pp. 459-471).

Rui, Y., Huang, T.S., & Chang, S.F. (1999). Image retrieval: Past, present, and future. *Journal of Visual Communication and Image Representation*, *10*, 1-23.

Smeulders, A.W.M., Gevers, T., & Kersten, M.L. (1998). Crossing the divide between computer vision and databases in search of image databases. *Proceedings of the Visual Database Systems Conference*, Italy (pp. 223-239).

White, D.A. & Jain, R. (1996). Similarity indexing with the SS-tree. *Proceedings of the 12th International Conference on Data Engineering*, New Orleans, Louisiana (pp. 516-523).

Yoshitaka, A. & Ichikawa, T. (1999). A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering, 11*(1), 81-93.

KEY TERMS

Directional Relation: Describes the order between two salient objects according to a direction, or the localization of salient object inside images. In the literature, 14 directional relations are considered:

- **Strict:** north, south, east, and west.
- **Mixture:** north-east, north-west, south-east, and south-west.
- **Positional:** left, right, up, down, front and behind.

Image Indexing: Consists of assigning concise and significant descriptors to an image. Objects shapes and positions are used in several indexing approaches where image is represented as a graph or tree (R-tree, B-tree, etc.).

Metric Relation: Measures the distance between salient objects. For instance, the metric relation "far" between two objects A and B indicates that each pair of points A_i and B_j has a distance grater than a certain value d.

Multimedia Document: Represents a document containing not only textual data but also multimedia ones such as images, videos, songs, and so on.

Salient Object: Interesting or significant object in an image (sun, mountain, boat, etc.). Its computing changes in function of the application domain.

Spatio-Temporal Relation: A combination of two spatial and temporal relations into one relation used to index video-documents and objects evolution between two images (e.g., tumor evolution).

Topological Relation: Describes the intersection and the incidence between objects. Six basic relations have been identified in the literature: *disjoint, meet, overlap, cover, contain, and equal*.

ENDNOTES

1 Where the interior set is empty, the boundary set is the barycentre, and the exterior is divided into four sets.

2 Traditionally, the three salient objects are described by non expressive spatial relations (same topological and directional relations).

3 Sequence of 2D images in medical applications (i.e., scanner).

The Elaboration Likelihood Model and Web-Based Persuasion

Kirk W. Duthler
The University of North Carolina at Charlotte, USA

INTRODUCTION

Discussing Google's reliance on the *AdWord* as a major source of revenue, *Wired's* Josh McHugh (2004) wrote of the obstacles faced by Google founders Sergey Brin and Larry Page in the late 1990's:

... the biggest challenge was convincing venture capitalists that Google could actually make money serving up minimalist, fast loading, text-only ads. It was 1998, after all, the heyday of elaborate splash pages and animated, brand-touting banners that danced at the top of every portal...Google didn't buy in—a stubbornness that proved brilliant. Six years later, those skinny little text-based ads are a huge money maker, accounting for more than $600 million in revenue last year... . (p. 120)

As McHugh (2004) points out, not only do persuasive appeals in digital media vary from the pallid and benign Google-like appeal to the flashy and vivid banner advertisement or corporate publicity site, but the simple Google appeal is highly successful. Recent research suggests a promising and powerful explanatory conceptualization of this continuum be based on a concept labeled peripheral cue complexity.

Peripheral cue complexity describes the degree to which a multimedia message contains production elements (visual and/or auditory effects), which are not directly related to the central meaning of the message. Rigorous experiment-based research reveals messages low in peripheral cue complexity, like that of the Google AdWord, are more appropriate and effective for highly involved and motivated individuals. While, messages with higher degrees of peripheral cue complexity pique the attention of minimally involved individuals and lead to more elaborate and focused cognitive processing of the message itself (Duthler, 2001; Singh & Dalal, 1999).

Peripheral cue complexity is derived from a significant theoretical model of persuasion called the *Elaboration Likelihood Model* (ELM) and extends the research into the cognitive processing of multimedia presentations. A significant body of research literature pertaining to the ELM from the social sciences of communication studies and social psychology helps the message designer and communication practitioner understand the information processing strategies of individuals faced with persuasive appeals. Recent incarnations of this literature may help to explain the wildly successful, yet plain Google AdWord. It may also explain the continued popularity of the Internet banner advertisements and the sophisticated, planned, visually complex corporate or commodity-related World Wide Web (WWW) site.

BACKGROUND

Though first proposed more than 20 years ago, the ELM (Petty & Cacioppo, 1981, 1986) helps to explain how information seekers process persuasive messages. Recent work (Duthler, 2001; Karsen & Korgaomkar, 2001; Singh & Dalal, 1999) to refine and adapt the ELM to digital media such as the Web has proven fruitful. An exploration of the fundamental tenants of the ELM, some criticisms, and recent refinements will help demonstrate its applicability to persuasion in digital media.

The ELM is an information processing theory of persuasion proposing two routes individuals take to analyze a persuasive appeal. The central route to persuasion, also labeled as central processing, involves high elaboration or careful scrutiny and thinking about an argument and its merits, to arrive at an evaluation of the advocated message. An individual taking the central route to persuasion carefully dissects the argument, weighing the data, arguments, and

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

warrants of the message. The central processor is one who pays ultimate attention to the informational content of the persuasive appeal. On the other hand, an individual taking the peripheral route to persuasion, also labeled peripheral processing, expends very little cognitive effort or low elaboration, instead relying on simple cues in the persuasive situation to arrive at an evaluation of a message. The peripheral processor foregoes consideration of the textual/informational dimension of the persuasive message, in favor of the sensory, non-content related dimension of the persuasive appeal.

According to the ELM, these routes to persuasion are assumed to be mediated by the motivation and/or ability of the individual. Because the central route is more difficult, a person with greater motivation is more likely to engage in central processing (Gass & Seiter, 2003). Motivation is typically operationalized by creating circumstances where outcome-relevant involvement is either high or low. Outcome-relevant involvement is the degree to which the economic or social outcome advocated in the message is important to the individual (Slater, 1997). When outcome-relevance is high, individuals are likely to take the central route. When outcome-relevance is low, individuals are likely to take the peripheral route. Even if an individual is highly motivated, they may not have the ability to process the message and thus must engage in peripheral processing. Ability can be affected by lack of previous knowledge, difficulty in analyzing complex material, distraction, a lack of time, or possibly a slow Internet connection. The key to the ELM is the proposal that when both motivation and ability are high, then elaboration likelihood is high and individuals are likely to follow the central route. However, when motivation and/or ability are low, elaboration likelihood is low and individuals are likely to follow the peripheral route. As either motivation or ability to process an argument are decreased, then peripheral cues become more important determinates of persuasion (Petty & Cacioppo, 1986).

Peripheral cues are variables that allow an individual to arrive at a judgment of an argument without processing the message arguments themselves (Petty & Cacioppo, 1986, p. 18). Commonly researched peripheral cues include source attractiveness (Forret & Turban, 1996), credibility or expertise (Petty, Cacioppo, & Goldman, 1981), argument length or number or arguments (Petty & Cacioppo, 1984), and even fragrances (DeBono, 1992). Furthermore, as individuals arrive at an attitude via the central route, attitudes are thought to be more accessible, persistent, resistant to change, and a better predictor of behavior than when the peripheral route is taken (Petty & Cacioppo, 1986). Research addressing the ELM is usually concerned with identifying the variables that affect elaboration likelihood (motivation and ability) and the effects of different variables (potential peripheral cues) in the persuasion context. An extensive research program supports these general relationships and conclusions.

Singh and Dalal's (1999) work is among the first published studies directly connecting the ELM with the WWW. The value of their study is the differentiation between the Web searcher and the Web surfer as central and peripheral processors, respectively. According to these researchers, the surfer is a hedonistic, fun-seeker and explorer who desires entertainment and stimulation…"likely to land at a Web site, linger for a brief period and take off for another more attractive site in their path" (p. 95). The surfer exemplifies the peripheral processor (low motivation/ability). The searcher is a goal-oriented, information seeker, likely to spend more time at preferred sites (p. 95). The searcher is typified by the central processor (high motivation and high ability).

Imagine Signh and Dalal's (1999) searcher attempting to find the best on-line value for a digital camera. Deciding to explore Froogle.com (Google's shopping site); the searcher types the model number of the digital camera into the search engine. Froogle.com returns eight AdWords and 21,200 total search results. The searcher not is likely to explore all 21,200 results, but will evaluate many WWW sites related to the search. The searcher will explore the primary search results and the AdWords, evaluating, price, retailer credibility, return policies, shipping prices, finally deciding on a retailer from whom to purchase the item. Contrast this to a surfer happening upon a manufacturer's Web site. Such intense comparison and evaluation will not take place. Rather the surfer might be drawn to the site because of the emotion-laden, eye-popping graphics or enticing interactivity. The surfer will spring for another site as soon as the initial interest is gone.

However, despite the explanatory power of the ELM it is not without its critics and has been reproached for a number of reasons. The foremost

criticism of the ELM is the lack of a sound, non-ambiguous, well-defined and concrete conceptualization of peripheral cues. According to Petty and Cacioppo (1986) "…peripheral cues refer to stimuli in the persuasion context that can affect attitudes without necessitating processing of the message arguments" (p. 18). Based on Petty and Cacioppo's definition or peripheral cues, one can conclude that peripheral cues are defined as any variable affecting attitudes in the absence of argument scrutiny. Such a definition of peripheral cues has been criticized for being ill-defined and ambiguous (Duthler, 2001; Eagly & Chaiken, 1993; Stephenson & Palmgreen, 2001). The central criticism of this attack is focused on Petty and Cacioppo's admission that a variable in the persuasion context can act in one of three roles. Petty and Cacioppo (1986) state that "variables can affect the amount and direction of attitude change by (a) serving as persuasive arguments, (b) serving as peripheral cues, and/or (c) affecting the extent or direction of issue and argument elaboration" (p. 16). Thus, source attractiveness, most commonly thought of as only a peripheral cue, may serve as a peripheral cue, an argument, or it may affect the extent or direction of message processing. Petty and Cacioppo have been severely criticized for this ambiguity and lack of operational precision (Duthler, 2001; Eagly & Chaiken, 1993; Stiff, 1986; Stiff & Boster, 1987; Stephenson, 1999; Stephenson & Palmgreen, 2001).

MAIN FOCUS OF THE ARTICLE

As a resolution to the debate concerning the ill-defined nature of peripheral cues, Stephenson and Palmgreen (2001) chose to focus more narrowly on non-argument-related cues important to a given medium. In their study televised public service announcements (PSAs), arguing against the use of marijuana, contained varying degrees of message sensation value. Message sensation value is defined as "the degree to which formal and content audio-visual features of a message elicit sensory, affective, and arousal responses" (p. 55). In that study, peripheral processing of messages was viewed as restricted to the sensory aspects of the PSAs, including sound and visual effects, music, and the use of close-ups, lighting, and camera angles. In this narrower conceptualization, peripheral cues are an important element in the production of messages in a given medium, but have little impact on the argumentative, information-laden content of the message. Peripheral cues are not operationalized as an attribute of a speaker or the number of arguments in a persuasive message or as any of the other typical peripheral cues.

Operationalizing a peripheral cue as an attribute of the source or the number of arguments potentially affects the strength or weakness of an argument (the content) and fundamentally changes the argument itself. However, operationalizing peripheral cues strictly as a property of a medium's production elements clearly separates them from the persuasive message. By re-conceptualizing peripheral cues in terms of the complexity of a message's production elements, the ambiguous and ill-defined nature of peripheral cues can be avoided or at least greatly reduced. Production elements on the Wevb can be easily manipulated to include varying degrees of color, animation, font type, non-content-related graphics, and even auditory cues. Variations in such elements are viewed as affecting the degree of peripheral cue complexity of a message.

Therefore, peripheral cue complexity describes a continuum ranging from relatively few production elements to a high degree of production elements. Peripheral cue complexity is a matter of degree, not an either or proposition. At one extreme is the Google AdWord and at the other is the highly stylized, programmed, complexity of the splash and flash commodity-related or corporate Web site. According to this conceptualization even the simple, text-based, Google AdWord contains at least some degree of production value. Each AdWord contains blue/magenta hyperlinked text of the sponsored key word, a black textual description, and the universal resource locator (Web address) of the sponsoring company emphasized in a light-green font color. In addition, there are two primary placements of the AdWord on Google's site: those shaded in blue at the top of the search results section of the page and those offset from the search results located in shaded area to the right of the search results. At the other end of the continuum is the commodity-related or corporate Web site. Contained in this site is a complex array of production elements including textual, musical, animated, visual, and other sophisticated production elements.

Duthler (2001) designed an experiment to test the effect of peripheral cue complexity within the context of the WWW on attitudes. A 2x2x2 factorial design manipulated peripheral cue complexity (high or low), outcome-relevant involvement (high or low), and argument strength (strong or weak). Manipulating outcome-relevance as high or low was similar to that of Singh and Dalal (1999) description of surfers (peripheral processors) and searchers (central processors). Peripheral cue complexity was operationalized by including or excluding pictures, color, icons, or clip art in specially designed WWW sites.

Hypotheses were consistent with the ELM, primarily focused on two predicted outcomes. First highly involved participants would pay little attention to peripheral cue complexity and process the message arguments. Second, low involvement participants would attend to peripheral cue complexity rather than message arguments.

Results of the experiment indicated that participants under conditions of high personal relevance (searchers) paid little attention to the peripheral cue complexity dimension of the Web sites, rather processing the claims, warrants, data, and conclusions of the arguments. As expected, the high personal relevance participants performed like searchers, rather than surfers, processing the messages centrally. Also expectedly, when personal relevance was low (the surfers), participants exposed to the low peripheral cue complexity sites hardly processed the information. However under the same surfer-like conditions (low personal relevance), participants browsing the high peripheral cue complexity Web sites unexpectedly processed the information much like that of a searcher.

Duthler (2001) concluded that higher degrees of peripheral cue complexity may induce central processing even among less involved individuals. Higher degrees of peripheral cue complexity grab the attention of the surfer, resulting in more searcher-like processing. It appears clear that peripheral cue complexity has significant explanatory qualities in predicting the behavior of internet surfers. In general, this study indicates that the degree of peripheral cue complexity may increase the attention paid to message content among those considered less personally involved in the persuasive outcome, and may have encouraged the processing of message arguments. Peripheral cue complexity does not seem to affect those considered involved in the message outcome.

CONCLUSION

As evidenced by Google's increasing dependence on the advertising revenue of the AdWord or the pervasiveness of highly sophisticated commodity-related Web sites such as Budweiser beer, Tide laundry detergent or any other highly produced corporate or product Web site and banner advertisements, the Web is a place infused with persuasive appeals. The ELM does well to explain these fundamentally different approaches to persuasion.

The ELM offers sophisticated explanations for Google's efficient, text-based, targeted advertisements resulting in clickthrough rates of 15%—a figure 10 times the effectiveness rate of banner advertising (McHugh, 2004). The low degree of peripheral cue complexity does not deter the central processor/searcher. In fact, such ads are precisely what the central processor is seeking—concise information directly related to the economic or social outcome sought allowing them to process significant amounts of information efficiently and thoroughly. On the other hand, the high degree of peripheral cue complexity designed into the banner ad or commodity-related WWW site is the perfect wave enticing the surfer.

FUTURE TRENDS

The future of research on peripheral cue complexity is bright. Research is needed to flesh out the effects of peripheral cue complexity. Such research should proceed in a number of directions. First, its relationship to the development, design, and technology of Web sites should be explored. Other independent variables may need to be included in studies related to peripheral cue complexity. Particularly within the domain of Internet-based media, download speed or bandwidth becomes an important variable to consider. As the degree of peripheral cue complexity increases so too does the time it takes to download the information to the recipients' computer. These delays between access and reception may significantly influ-

ence the effectiveness of persuasive communications. This raises the question of how much is too much? In other words, at what point does the degree of peripheral cue complexity become detrimental, resulting in distraction, less attention to arguments, and perhaps less persuasion?

REFERENCES

DeBono, K. (1992). Pleasant scents and persuasion: An information processing approach. *Journal of Psychology, 102*, 91-102.

Duthler, K. (2001). *The effect of peripheral cues on the processing persuasive messages on the World Wide* Web. Unpublished Doctoral Dissertation, University of Kentucky, USA.

Eagly, A. & Chaiken, S. (1993). *The psychology of attitudes.* Fort Worth, TX: Harcourt Brace College Publishers.

Forret, M. & Turban, D. (1996). Implications of the elaboration likelihood model for the interviewer decision process. *Journal of Business and Psychology, 10*(4), 415-429.

Gass, R. & Seiter, J. (2003). *Persuasion, social influence, and compliance gaining* (2nd ed.). Boston: Allyn & Bacon.

Karson, E. & Korgaonkar, P. (2001). An experimental investigation of Internet advertising and the elaboration likelihood model. *Journal of Current Issues and Research in Advertising, 23*(1), 53-72.

McHugh, J. (2004). It's an ad, ad, ad, ad, ad world. *Wired*, *12*(3), 120-121.

Petty, R. & Cacioppo, J. (1981). *Attitudes and persuasion: Classic and contemporary approaches.* Dubuque, IA: William C. Brown.

Petty, R. & Cacioppo, J. (1984). Effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology, 46*, 69-81.

Petty, R. & Cacioppo, J. (1986a). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

Petty, R. & Cacioppo, J. (1986b). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology,* (Vol. 19). New York: Academic Press.

Petty, R., Cacioppo, J., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology, 41,* 847-855.

Petty, R., Cacioppo, J., Kasmer, J., Haugtvedt, C. & Cacioppo, J. (1987a). A reply to Stiff and Boster. *Communication Monographs, 54,* 258-263.

Petty, R., Kasmer, J., Haugtvedt, C., & Cacioppo, J. (1987b). Source and message factors in persuasion: A reply to Stiff's critique of the elaboration likelihood model. *Communication Monographs, 54,* 233.

Singh, S. & Dalal, N. (1999). Web home pages as advertisements. *Communications of the ACM*, *42*(8), 91-98.

Slater, M. (1997). Persuasion processes across receiver goals and message genres. *Communication Theory, 7*(2), 125-148.

Stephenson, M.T. (1999). *Message sensation value and sensation seeking as determinants of message processing*. Unpublished Doctoral Dissertation, University of Kentucky, USA.

Stephenson, M.T. & Palmgreen, P. (2001). Sensation seeking, message sensation value, personal involvement, and processing of anti-drug PSAs. *Communication Monographs*, *68*(1), 49-72.

Stiff, J. (1986). Cognitive processing of persuasive message cues: A meta-analytic review of the effects of supporting information on attitudes. *Communication Monographs, 53,* 75-89.

Stiff, J. & Boster, F. (1987). Cognitive processing: Additional thoughts and a reply to Petty, Kasmer, Haugtvedt, and Cacioppo. *Communication Monographs, 54,* 250-256.

KEY TERMS

Central Route to Persuasion: A term used in the elaboration likelihood model involving intense thought and analysis concerning a persuasive message. The central route to persuasion involves high degrees of message related thinking or careful scrutiny about an argument and its merits in order to arrive at an evaluation of the advocated message. When the central route to persuasion is taken, attitudes are thought to be more accessible, persistent, resistant to change, and better predictors of behavior than when the peripheral route is taken.

Elaboration: Issue or message relevant cognition. Elaboration is typically conceptualized as a continuum ranging from high elaboration to low elaboration. It is thought to be influenced by an individual's motivation and ability to cognitively process a message.

Elaboration Likelihood Model: The ELM is an information processing theory of persuasion first proposed in 1981 by social psychologists Richard Petty and John Cacioppo. The model of persuasion proposes individuals either think carefully about the acceptability of a persuasive appeal or instead with little cognitive effort rely on cues in the persuasive situation to arrive at a conclusion of the advocacy.

Google AdWords: A plain, text-based advertisement displayed prominently on Google.com. For a fee, advertisers sponsor search words/terms entered by users of the Google.com search engine. The sponsored search words are predicted by the sponsor to relate to a product or service offered.

Message Sensation Value: The degree to which formal and content audio-visual features of a message elicit sensory, affective, and arousal responses (Stephenson & Palmgreen, 2001, p. 5).

Outcome-Relevant Involvement: The degree to which the economic or social outcome advocated in the message is important to the individual (Slater, 1997).

Peripheral Cue: A variable in the persuasion situation that allows an individual to arrive at a judgment of an argument without processing the message arguments themselves (Petty & Cacioppo, 1986, p. 18).

Peripheral Cue Complexity: A term describing the degree to which a multimedia message contains production elements (visual and/or auditory effects), which are not directly related to the central meaning of the message. Peripheral cue complexity is envisioned as a continuum anchored by low peripheral cue complexity at one end and high peripheral cue complexity at the other.

Peripheral Route to Persuasion: A term used in the elaboration likelihood model involving a lack of intense thought and analysis concerning a persuasive message. An individual taking the peripheral route to persuasion relies on simple cues in the persuasion context to arrive at an evaluation of the advocated message.

E-Learning and Multimedia Databases

Theresa M. Vitolo
Gannon University, USA

Shashidhar Panjala
Gannon University, USA

Jeremy C. Cannell
Gannon University, USA

INTRODUCTION

E-learning covers the variety of teaching and learning approaches, methodologies and technologies supporting synchronous or asynchronous distance education. While distance education is a concept typically used by conventional institutions of education to mean remote access and delivery of instruction, the concept of e-learning broadens the scope to all instances of learning using Web-mediated learning. The scope includes realizing learning organizations (Garvin, 1993), achieving knowledge management (Beccerra-Fernandez; Gonzalez & Sabherwal, 2004; Aussenhofer, 2002) and implementing organizational training.

Individuals continue to learn throughout their lives, particularly as a function of their work and profession. The manner in which they access information and use it often depends upon the available technology, their previously learned response for information acquisition and how their organization facilitates learning and knowledge transfer (Tapscott, 1998; Zemke, Raines & Filipczak, 2000). Hence, e-learning is not simply a consideration for traditional learning institutions, but for any organization.

As such, e-learning not only faces the traditional challenges of teaching to various learning styles while conveying the spectrum of educational objectives, but also faces the extra challenge of using emerging technologies effectively. The three significant emerging technology areas to e-learning are: networking, mobility and multimedia. These technologies can enable a highly interactive delivery of material and communication between instructors and students. Out of the three, however, multimedia technologies relate directly to pedagogical concerns in providing material tailored to the content domain, to the individual and to the learning objectives (Vitolo, 1993).

Currently, multimedia and e-learning initiatives focus on the presentation of multimedia. The adequate presentation of multimedia is often more an issue of the network being used and its connectivity parameters. Acceptable multimedia presentation depends upon the format of the multimedia and its ability to be quickly transferred (David, 1997). In these circumstances, the availability and appropriateness of the multimedia is assumed to have already been decided as necessary to the instruction.

Not being addressed currently is the storage of multimedia. Multimedia databases should allow for retrieval of components of the integrated and layered elements of the media data stored. In this way, the media would support learning goals. Its retrieval should be conditional upon a context and a content need. Context involves the learning situation – the educational objectives and the learner, combined. Content need includes the particular material to be acquired. Conditional retrieval of multimedia based upon a pedagogical circumstance implies that not all learners or situations need the same media to be delivered, but that a compendium of stored media should be available. In fact, the media alone cannot solely enable learning. Clark (1983) analyzed the effects of learning from different media and observed that significant changes in learning are a function of the media used for the presentation of the material. Significant attention must be given to the content material available for e-learning systems. The material in a certain media format should be included, because it adds or complements the underlying informational intent of the system.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Further, as educational objectives aspire to higher levels of competency such as analysis, synthesis and evaluation, more depth and variety of detail need to be communicated to the student. However, due to the connectivity issues of e-learning, often layers of representation are not available to the learner. For example, during face-to-face communication, student to teacher, the teacher provides the path to the solution and essentially trains the student when teaching analysis skills. However, with e-learning systems, just the end product—the "solution"—of the analysis is provided. When the underlying reasoning layers of the analysis are not available, the overall quality of the instruction suffers (Vitolo, 2003).

Multimedia databases added to an e-learning initiative would provide conditional retrieval and comprehensive storage of multimedia. However, no database management system (DBMS) exists solely for multimedia storage and access (Elmasri & Navathe, 2000). Several current DBMS do provide a data type appropriate for multimedia objects. However, the range of capabilities available for manipulating the stored object is severely limited. A pure multimedia database management system (MDDBMS) is not commercially available, now.

BACKGROUND

Learning, education and teaching are inextricably intertwined, highly complex processes. Each process has been researched as a social phenomenon, cognitive transformation, generational bias and personality expression. While the work on these topics is vast, several aspects are generally accepted as foundation concepts:

- People interact with environments on an individualized basis. Learners have learning styles; teachers have teaching styles; individuals have personality styles.
- Educational efforts seek to find a correspondence between these various styles so that learning can progress effectively.
- Educational efforts can be described via taxonomies—progressions of objectives. The realization of these objectives does not necessarily require any specific learning or teaching modality. The communication of the content of the objective may be better suited to one modality (visual, auditory or tactile) than another.
- Learning can continue throughout an individual's life.
- Technology can facilitate educational efforts by providing various formatted and comprehensive content for interactive and self-regulated learning. Multimedia technology provides an excellent opportunity for packaging content into a variety of modalities.

With respect to styles, Coates (2002) provides a condensation of the various style-based perspectives of learning. While much of these style-based analyses of behavior stem from the initial work of Carl Jung (1923), the facets of the styles are continually being researched. Learning is mediated by a variety of factors—some (such as modality of instruction) that can be manipulated successfully within an educational effort, some (such as generational cohort biases) that are out of the control of instructional design.

With respect to educational structures, educational researchers have developed taxonomies to explain educational objectives. (See Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths & Wittrock, (2001) and Bloom (1984, 1956) for classic coverage of these taxonomies.) Essentially, educational efforts advance instruction in levels of difficulty and performance so that the breadth and depth of the knowledge of a field can be communicated.

As a foundation concept to using multimedia for e-learning, the media requires appropriate processing for adequate capture, production and distribution. For example, video may be shot using either an analog or digital camera. Before the source video can be edited using computer software, it must be instantly accessible from a hard disk and not the original videotape. The source video is imported into the computer by a process called video capture. Captured video is huge; 10 seconds of raw, uncompressed NTSC video (the standard for television video) use as much as 300 megabytes (MB) of storage space.

For video to be played in a Web browser or distributed on CD-ROM, the file size must be reduced significantly. This file size reduction, or com-

pression, is achieved using codecs – compression/decompression approaches. Source video captured from a digital camcorder will already have been digitized and saved in a digital file format inside the camera. Digitizing a video sequence results in extremely high data rates. For example, an image with a resolution of 720x576 pixels and a color depth of 16 bits produces a data stream of 1.35 MB per individual frame. At the rate of 25 frames per second required to render smooth video scenes, a gigantic data volume of 3,375 MB/second results. This volume is far too great for the average hard disk to handle; a CD-ROM would only have enough space for about 16 seconds (Adobe Press, 2003; Bolante, 2004).

Next, the capture process involves transferring the digitized video file to a computer hard disk. Once captured, the multimedia requires further considerations for production and dissemination considerations. The analog or digital source video is captured using video editing software and saved into an appropriate video format. These video formatted files are also large; 60 minutes of video can consume 12 GB of disk space. The media file is manipulated within software via timing option, making it ready for rendering and production. After rendering, the video file is processed further depending upon its desired distribution modality:

- Exported back to video tape (analog or digital)
- Compressed further for distribution on CD-ROM or DVD
- Compressed further for distribution across the Internet

The final presentation also has options. Progressive encoding refers to where the entire video must be downloaded before any viewing occurs, regardless of its format. This case occurs with any of the formats considered so far. Alternatively, Internet streaming enables the viewer to watch sections of video without downloading the entire file. Here, the video starts after just a few seconds. The quality of streaming formats is significantly lower than progressive formats due to the compression being used (Menin, 2002).

Finally, appropriate display of the material for effective consumption is improved with interactive multimedia. However, interaction with a media file—the goal of interactive multimedia—is restricted; navigation is possible using pause, forward and reverse controls provided by the player installed on the client computer. To create interactive media for the Web, CDs, kiosks, presentations and corporate intranets, a multimedia authoring program is used. These programs enable the combination of text, graphics, sound, video or vector graphics in any sequence. To add more interactive features, powerful scripting languages are also provided (Gross, 2003).

Hence, the situation for e-learning is bound in several ways by the available multimedia technology. First, the production and distribution of multimedia is not a trivial undertaking, requiring specialized skills and technologies. Second, the viewing of the multimedia requires the client machine and user to have appropriate technology. Third, the goals of the e-learning effort must be in balance with the available and expected technology. Fourth, the multimedia technology itself is providing limited options for interactive manipulations. After these steps, the media as a data-rich structure can be stored in multimedia databases.

LIMITED STATE OF E-LEARNING AND MULTIMEDIA DATABASE SYSTEMS

The requirements for the next era of e-learning applications using multimedia databases are:

1. Repository systems offering storage and access capabilities of media
2. Indexable storage structure for media files as contiguous structures composed of identifiable and searchable elements
3. Tier-architecture deployment providing multiple application access

With respect to point 1, DBMS implementations are commercially available that can reference media files in a variety of formats. The media file is handled as a complete unit through a large object (LOB) data type. For instance, Oracle introduced a set of LOB data types with Oracle 8 to facilitate the storage of large-scale digitized structures and references to them. The media itself is stored in one of two ways: as a LOB (usually BLOB) type within the database, or in an external file and pointed to by a BFILE type within the database.

Oracle has continued to advance the integration of these LOB data types through its database versions and the various tools it offers. Oracle's *inter*Media management system and Oracle 10g database support various media specific object types, recognize and record facets of the media's attributes in metadata structures, and provide support to multimedia needs for various applications and enterprise-wide delivery. The *inter*Media objects of ORDAudio, ORDImage, ORDVideo and ORDDoc provide attributes and access capabilities recognized by the end-user application (Oracle, 2003, 1999).

With respect to point 2, however, media data are not handled as an indexed structure. Thus, access to incremental slices or partitions of the media is not a current capability. Access to the attributes of the entire media as a unit has been improved, but more is desired for e-learning. A beginning point would be a query standard for multimedia and its internal elements. Oracle 10g does support a portion of the first edition of the ISO/IEC 13249-5:2001SQL/MM Part 5: Still Image Standard. The standards community and commercial vendors continue to address the query needs of multimedia and applications.

With respect to point 3, multimedia applications may not have any database component. The media is handled as a data item manipulated by the script of the application (David, 1997). This situation mimics the file-processing era; namely, that the data manipulated by one file for its application's needs may not lend itself to manipulation for another, separate application's needs.

A more desirable architecture for multimedia delivery is a three-tier one. The advantages of a multi-tier architecture are:

- Separation of the user interface logic and business logic
- Low bandwidth network requirements
- Business logic resides on a small number (possibly only one) of the middle machines

Together, these aspects promote greater accessibility across various applications and ease of maintenance through hardware and software upgrades.

E-learning has benefited from the enhanced broadband and accessibility of the public infrastructure. As such, the popularity and market presence of the initiatives have grown. The initiatives offer delivery convenience, communication channels and content volume. These three factors make e-learning a highly attractive possibility for a variety of learning circumstances.

In many respects, however, current e-learning initiatives are similar to page-turning, computer-assisted instruction packages of earlier computer-based learning efforts. That is, the initiatives lack pedagogical development and refinements, tailoring the instruction with respect to a student model and to the complexities of higher-level educational objectives.

The higher levels of educational objectives need to communicate greater complexity in the detail, explanation and incremental refinements of the content. If only the final result of the analysis, synthesis or evaluation is the goal of the instruction, then current multimedia efforts would be fine. However, higher-level educational objectives require more layers and connections to be communicated—more complex structures need to be communicated in order to teach more complex concepts. Further, this type of refinement of complex structures through incremental layers of development, feedback, and progress is desirable for effective education and for effective handling of learning styles given a learning situation. To achieve mediated learning episodes with this finesse, the next generation of multimedia applications and deployment utilities are necessary.

FUTURE TRENDS & CHALLENGES

Multimedia capabilities will continue to improve, becoming more economical and more usable. In time, the authoring, production and distribution of multimedia will become as easy as word processing. As with many information systems efforts, the challenge resides with understanding the infrastructure commitment to deploy such efforts in terms of hardware, skills and procedures. Successful efforts require high-capacity, secure servers and connections. Individuals need to understand the nature of multimedia to manipulate it successfully within the software. Finally, well-defined procedures for the distribution and maintenance of the multimedia over

a desired architecture must accompany the effort and must be handled by systems staff cognizant of the desired performance levels.

The future of multimedia databases shares this same positive outlook. The capabilities sought in various data-typing of media, querying of media segments and indexed aspects need continued addressing.

For e-learning, the challenges parallel those of multimedia. E-learning efforts need to understand how the infrastructure can limit the instructional goals. The skill level for development efforts requires technical competence and instructional design principles. When future e-learning efforts include multimedia databases, then the required technical skills will be further specialized. E-learning efforts will require teams of highly specialized individuals, bridging the different technical needs for pedagogy, multimedia and multimedia databases.

The final future challenge to be addressed is one shared with many Web-based developments—intellectual property rights. Intellectual property is a sufficiently difficult concept currently when multimedia is part of a single application. Once the multimedia is part of applications connected through a database, then the intellectual property rights of the database and its development must be considered also.

CONCLUSION

E-learning continues the efforts of computer-mediated instruction. The depth and interactivity potential of multimedia components is a highly attractive factor to add to instruction. Multimedia offers the capability to construct interactions tailored to the learning needs of a specific student, within a specific learning context, being taught a specific content domain.

Multimedia technology has matured significantly as its complementary technologies of network capacities and deployment hardware have advanced. However, for the next generation of multimedia and e-learning to progress, multimedia databases should be used. The database configuration would increase the potential use of the multimedia across multiple application instances, the multimedia could be queried for access and, ultimately, the elements composing the multimedia could be accessed as opposed to accessing the entire multimedia file—the current option for multimedia access.

Multimedia databases not only would enhance the technical delivery of e-learning efforts, but also would enhance the pedagogical aims of e-learning efforts. Instruction of the higher-order educational objectives requires layers of a representation to be presented. Multimedia databases could store media in its elemental segments so that selective delivery of pieces of the media could be offered for instruction—not the media file in its entirety, leaving the parsing of the relevancy of the media to the discretion of the student.

While multimedia databases would increase the flexibility, access and reuse of the media, other challenges arise. Multimedia databases are complex technologies requiring more specialized skills beyond simply building and deploying multimedia. Adequately supporting multimedia databases requires continued, expensive investments in infrastructure to support the deployment of the databases and e-learning efforts. Further, not all of the required features of true multimedia databases have been developed to date, but are part of the current efforts of database developers and of standards communities. Finally, intellectual property issues are a challenge of applications using multimedia databases. As in most development aspects, the intellectual property issues will be as difficult to resolve as the technology was to develop.

REFERENCES

Adobe Press. (2003). *Adobe Premiere 6.5: Classroom in a book.* San Jose: Adobe Systems International.

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *A taxonomy for learning, teaching, and assessing*. New York: Longman Publishers.

Ausserhofer, A. (2002). E-learning & knowledge management towards life-long education. Graz, Austria: Competence Center for Knowledge-based

Applications and Systems. Retrieved December 10, 2003, from *www.know-center.tugraz.at/de/divisions/publications/pdf/aausser2002-01.pdf*

Becerra-Fernandez, I., Gonzalez, A., & Sabherwal, R. (2004). *Knowledge management: Challenges, solutions, and technologies*. Upper Saddle River, NJ: Pearson Education.

Bloom, B.S. (Ed.) (1984, 1956). *Taxonomy of educational objectives. Handbook 1: Cognitive domain*. New York: Longman.

Bolante, A. (2004). *Premiere Pro for Windows.* San Jose: Peachpit Press.

Clark, R.E. (1983). Reconsidering research on learning from media. *Review of Educational Research*, *53*(4), 445-459.

Coates, J. (2002). *Generational learning style*. Retrieved May 23, 2002, from *www.lern.org/gls_booklet.pdf*

David, M.M. (1997). Multimedia databases through the looking glass. *Intelligent Enterprise's Database Programming & Design: On-Line.* Retrieved December 10, 2003, from *www.dbpd.com/ vault/ 9705davd.htm*

Garvin, D.A. (1993). Building a learning organization. *Harvard Business Review*, *71*(4), 78-92.

Gross, P. (2003). *Macromedia Director MX and lingo: Training from the source*. Berkeley, CA: Macromedia Press.

Jung, C. (1923). *Psychological types.* New York: Harcourt & Brace.

Menin, E. (2002). *The streaming media handbook.* Upper Saddle River, NJ: Prentice Hall.

Navathe, S.B., & Elmasri, R. (2000). *Fundamentals of database systems* (3rd ed.). Reading, MA: Addison-Wesley.

Oracle. (1999). *Using Oracle 8i* inter*Media with the Web, Release 8.1.5.2* (Part No. A77033-01). Retrieved May 20, 2004.

Oracle. (2003). *Oracle* inter*Media: Managing Multimedia Content* (Oracle white paper). Retrieved May 20, 2004.

Tapscott, D. (1998). *Growing up digital: The rise of the Net generation*. New York: McGraw-Hill.

Vitolo, T.M. (1993). The case for self-revealing multimedia systems. *Proceedings of the 11th Annual Conference of the Association of Management,* Atlanta, Georgia, August 5-9.

Vitolo, T.M. (2003). *The importance of the path not taken: The value of sharing process as well as product. Final report*. Vancouver: SMARTer Kids Foundation.

Zemke, R., Raines, C., & Filipczak, B. (2000). *Generations at work: Managing the clash of veterans, boomers, Xers, and nexters in your workplace*. New York: AMACOM.

KEY TERMS

Bit Depth: The number of bits used for color resolution when viewing a movie.

Codec: Compression and decompression algorithms provided by either a software application or a hardware device.

Database Management System (DBMS): Collection of software components to store data, access the data, define data elements, store data element definitions, build data storage structures, query the data, backup and secure the data, and provide reports of the data.

E-Learning: All teaching and learning processes and functions from course authoring, course management, examinations, content delivery, feedback and course administration developed, delivered and monitored through synchronous or asynchronous communication.

Encoding: The process of using codecs to convert video files to different distribution file formats. The codecs used for encoding files for CD-ROM and DVD are MPEG-1 and MPEG-2, respectively.

Frame Rate: The number of frames projected per second.

Frame size: The height and width of the video window according to the number of pixels.

Internet Streaming: Video format that intermittently downloads sections of a media file to a client.

Knowledge Management: The set of initiatives to identifying, retrieving, organizing, disseminating and leveraging intellectual capital—usually as an enterprise-wide effort.

Learning Styles: A cognitive perspective of individualized preferences for modalities when learning; includes a learner's manner of responding to and using stimuli while learning.

Multimedia Database: Database storage and retrieval capabilities developed with respect to multimedia requirements for high-quality, rapid, queried usage by applications.

Progressive Encoding: Video format that downloads the entire media file to the client before any displaying occurs.

Electronic Commerce Technologies Management

Shawren Singh
University of South Africa, South Africa

ELECTRONIC COMMERCE

The first e-commerce (EC) applications were started 30 years ago, in the early 1970s. The original applications were in the form of electronic fund transfers (EFT). These applications were limited to larger corporations and financial institutions (Turban, Lee, King & Chung, 2000). This type of transaction later included electronic data interchange (EDI). There is a marked difference between EDI and EC in that EC involves much more than EDI (Greenstein & Feinman, 2000).

There is no standard definition for EC, although a number of researchers have attempted to define it (Greenstein & Feinman, 2000; U.S. Department of Commerce, 1999). In principle, however, most authors are in agreement that EC uses some form of transmission medium through which exchange of information takes place in order to conduct business (Barnard & Wesson, 2000).

TYPES OF EC

There are different types of classifications of EC; for instance, by participants, task and technology.

Classification According to Participants

Turban, King, Lee, Warkentin and Chan (2002) provide the following definitions for the different types of EC, together with additional types that researchers have identified:

- **Business-to-Business (B2B):** This includes inter-organizational information systems and electronic transactions between organizations. An example of B2B is General Electric's Trading Process Network (TPN)(www.tpn.geis.com).
- **Business-to-Consumer (B2C):** B2C transactions are mostly retailing transactions with individual customers or consumers. An example of B2C is Amazon.com (www.amazon.com).
- **Consumer-to-Business (C2B):** In this category one will find consumers who sell to organizations. It also includes individuals who seek sellers with whom they may interact in order to conclude a transaction. An example of C2B is Priceline (www.priceline.com).
- **Consumer-to-Consumer (C2C):** C2C involves consumers selling directly to other consumers. This type of application includes auction sites and advertising personal services on the Internet. It can also include intranets and other organizational networks to advertise items and services. An example of C2C is eBay (www.eBay.com).

The additional types of EC identified by the above researchers are:

- **People-to-people (P2P):** This type of transaction is a special type of C2C where people exchange CDs, videos, software and other goods (www.napster.com).
- **Non-business EC:** Many institutions or organizations also use EC to improve their operation and customer services.
- **Intrabusiness (organizational) EC:** All internal organizational activities involving exchange of goods, services or information usually performed on intranets are included in this category.
- **Business-to-employees (B2E):** This is a subset of the intrabusiness category, where the organization delivers services, information or products to individual employees.
- **Government-to-citizen (G2C):** and to others: In this type of EC, a government entity buys or sells goods, services or information to businesses or individual citizens.
- **Exchange-to-exchange (E2E):** With the proliferation of exchanges and portals, it is logical for

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

exchanges to connect to one another. E2E is a formal system that connects exchange.

- **Collaborative commerce (c-commerce):** C-commerce is an application of an interorganizational information system for electronic collaboration between business partners and organizational employees.
- **Ultimate commerce (u-commerce):** U-commerce is the use of ubiquitous networks to support personalized and uninterrupted communications and transactions between a firm and its various stakeholders to provide a level of value over, above and beyond traditional commerce (Watson, 2000).
- **Mobile commerce (m-commerce):** When EC takes place in a wireless environment.

Classifications According to Task

EC can also be classified by the nature of the task. There are many different types of EC, such as e-shopping, e-banking and e-investments.

E-Shopping

Among the different EC activities, an e-shopping task has the following two unique phases: the look-see-and-decide phase (LSD), and the checkout phase (Renaud, Kotze, & van Dyk, 2001), as depicted in Figure 1.1. The LSD allows the user to browse, while a commitment to buy takes place in the checkout phase.

- **LSD:** This stage typically will be used to look at available products, compare them and then make a decision about whether to purchase products. This may be done once or more, often until the consumers have found products that satisfy their needs. This phase is intensely user-driven, because the user is looking at and assimilating information continuously. It has the following substages, which can be traversed iteratively and in varying sequences: welcome, search, browse and choose.
- **Checkout:** When users trigger this stage, they have made their choice of offered products and have decided to make a purchase. They now have to provide certain details, such as their address and credit card details. This stage is system-driven and changes the paradigm of the interaction process from user initiative to system initiative. Feedback is of critical importance during this stage—users who feel that they have lost control can simply leave the site without any embarrassment—unlike a user who is standing at a checkout in a supermarket. This stage is typically composed of at least the following steps, which should be navigated in a serial fashion: identifying the user, where the delivery should take place, how it should take place, payment, confirmation of order and completion (closure).

E-Banking

Internet-based services allow banking clients to obtain account information and balance enquiries; execute account payments and inter-account fund transfers; make queries on account balances; obtain statements; and, in some cases, view images of checks (Chan et al., 2001) from the comfort of their homes or offices. Additionally, by linking their accounts to personal finance software (such as Intuit Quicken and Microsoft Money), they will be able to track their spending offline, and later reconcile that with their bank statements online.

A typical e-bank task would be: Launch browser => Go to this bank's page: (URL to bank; e.g., www.absa.co.za) => Locate Internet banking and click that option => Login using this account number and password: (account_number), (password) => Choose type of transaction to conduct (this could consist of several steps)(user may choose to conduct more than one transaction) =>Logoff from bank's Web site => Close browser.

E-Investments

E-investments is a process that allows a user to trade stocks, bonds, mutual funds and other financial equities on the Internet. These companies offer users the opportunity to trade at a very small cost compared to discount brokers or full-service brokers. This has resulted in online trading companies grabbing an increasing market share (Chan et al., 2001).

A typical e-investment task to purchase stocks would be: Launch browser => Go to this broker's

page: (URL to bank; e.g., www.Datek.com) => Locate trade stocks and click that option => Login using this account number and password: (account_number), (password) => Choose type of transaction to conduct (this could consist of several steps)(user may choose to conduct more than one transaction) => Compare products (product A) and (product B). Determine which product has the highest performance in terms of (key product performance dimension) => Purchase chosen product => Confirm purchase => Logoff from Web site => Close browser

Figure 1 is but a small sample of the e-revolution. There are other e-activities, such as: e-tailers, e-insurance, e-travel, e-consulting, e-training, e-support, e-recruitment, all the way to e-cooking!

Classification with Technology

The Internet economy can be conceptualized as a collection of IP-based networks, software applications and the human capital that makes the networks and applications work together for online business, and agents (corporations and individuals) who are involved in buying and selling products and services in direct and indirect ways. There is a natural structure or hierarchy to the Internet economy that can be traced to how businesses generate revenue. Based upon this type of structure, Whinston, Barua, Shutter, Wilson and Pinnell (2000) broadly classify the Internet economy into infrastructure and economic activity categories, as seen in Figure 2.

The infrastructure categories are further divided into two distinct but complementary "layers": the Internet infrastructure layer, which provides the physical infrastructure for EC, and the Internet application infrastructure, which includes software applications, consulting, training and integration services that build on top of the network infrastructure, and which make it feasible for organizations to engage in online commerce.

The economic activity category is also subdivided into two layers: electronic intermediaries and online transactions. The intermediary layer involves the role of a third party in a variety of capacities: market maker, provider of expertise or certification that makes it easier for buyers to choose sellers and/or products, search and retrieval services that reduce transaction costs in an electronic market, and other services that facilitate conducting online commerce. The transactions layer involves direct transactions between buyers and sellers like manufacturers and e-tailers.

Layer One: The Internet Infrastructure Indicator

The Internet infrastructure layer includes companies that manufacture or provide products and services that make up the Internet network infrastructure. This layer includes companies that provide telecommunications and fiber backbones, access and end-user networking equipment necessary for the proliferation of EC. This layer includes the following types of companies: National and regional backbone providers (e.g., Qwest, MCI WorldCom); Internet Service Providers (e.g., AOL, Earthlink); network equipment for backbones and service providers (e.g., Cisco, Lucent, 3Com); conduit manufacturers (e.g., Corning); and server and client hardware (e.g., Dell, Compaq, HP).

Figure 1.1. LSD model

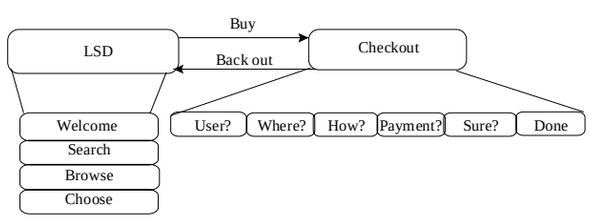
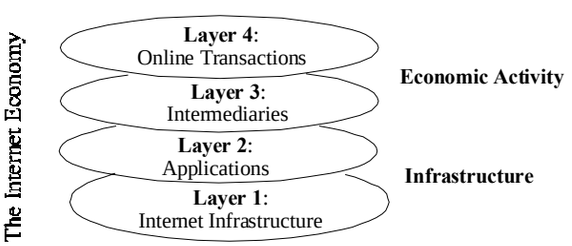


Figure 2. Classification of the Internet economy



Layer Two: The Internet Applications Infrastructure Layer

Products and services in this layer build upon the above IP network infrastructure and make it technologically feasible to perform business activities online. In addition to software applications, this layer includes the human capital involved in the deployment of EC applications. For example, Web design, Web consulting and Web integration are considered to be part of this layer. This layer includes the following categories: Internet consultants (e.g., MarchFIRST, Scient); Internet commerce applications (e.g., Microsoft, Sun, IBM); multimedia applications (e.g., RealNetworks, Macromedia); Web development software (e.g., Adobe, Allaire, Vignette); search engine software (e.g., Inktomi, Verity); online training (e.g., Sylvan Prometric, SmartPlanet); Web-enabled databases; network operating systems; Web hosting and support services; transaction processing companies.

Layer Three: The Internet Intermediary Indicator

Internet intermediaries increase the efficiency of electronic markets by facilitating the meeting and interaction of buyers and sellers over the Internet. They act as catalysts in the process through which investments in the infrastructure and applications layers are transformed into business transactions.

Internet intermediaries play a critical role in filling information and knowledge gaps, which would otherwise impair the functioning of the Internet as a business channel. This layer includes: market makers in vertical industries (e.g., VerticalNet, PCOrder); online travel agencies (e.g., TravelWeb, Travelocity); online brokerages (e.g., E*trade, Schwab.com, DLJ direct); content aggregators (e.g., Cnet, Cdnet); portals/content providers (e.g., Yahoo, Excite); Internet ad brokers (e.g., DoubleClick, 24/7 Media); online advertising (e.g., Yahoo, ESPN Sportszone); Web-based virtual malls (e.g., Lycos shopping).

Layer Four: The Internet Commerce Indicator

This layer includes companies that generate product and service sales to consumers or businesses over the Internet. This indicator includes online retailing and other business-to-business and business-to-consumer transactions conducted on the Internet. This layer includes: e-tailers selling books, music, apparel, flowers and so forth over the Web (e.g., Amazon.com, 1-800-flowers.com); manufacturers selling products direct such as computer hardware and software (e.g., Cisco, Dell, IBM); transportation service providers selling tickets over the Web (e.g., Delta, United, Southwest); online entertainment and professional services (e.g., ESPN Sportszone, guru.com); shipping services (e.g., UPS, FedEx).

It is important to note that many companies operate in multiple layers. For instance, Microsoft and IBM are important players in the Internet infrastructure, applications and Internet commerce layers, while AOL/Netscape has businesses that fall into all four layers. Similarly, Cisco and Dell are important players in both the infrastructure and commerce layers.

Each layer of the Internet economy is critically dependent on every other layer. For instance, improvements in layer one can help all the other layers in different ways. As the IP network infrastructure turns to broadband technologies, applications vendors in layer two can create multimedia applications that can benefit from the availability of high bandwidth.

CONCLUSION

Understanding the classification of one's e-activity could possibly improve a company's strategic/competitive edge in the market.

REFERENCES

Barnard, L., & Wesson, J. (2000, November 1-3). *E-commerce: An investigation into usability issues.* Paper presented at the 2000 South African Institute of Computer Scientists and Information Technologists (SAICSIT), South Africa, Cape Town.

Chan, H., Lee, R., Dillon, T., & Chang, E. (2001). *E-commerce: Fundamentals and applications*. Chichester, UK: John Wiley & Sons.

Greenstein, M., & Feinman, T.M. (2000). *Electronic commerce: Security, risk management and control*. Boston: Irwin McGraw-Hill.

Renaud, K., Kotze, P., & van Dyk, T. (2001). A mechanism for evaluating feedback of e-commerce sites. In B. Schmid, Stanoevska-Slabeva & V. Tschammer (Eds.), *Towards the e-society: E-commerce, e-business, and e-government* (pp. 389-398). Boston: Kluwer Academic Publishers.

Turban, E., King, D., Lee, J., Warkentin, M., & Chan, H.M. (2002). *Electronic commerce 2002: A managerial perspective* (2nd ed.). NJ: Prentice Hall.

Turban, E., Lee, J., King, D., & Chung, M.H. (2000). *Electronic commerce: A managerial perspective*. NJ: Prentice Hall.

U.S. Department of Commerce. (1999). The emerging digital economy II. Retrieved March 1, 2000, from *www.ecoomerce.gov/edu/chapter1.html*

Watson, R.T. (2000). U-commerce - The ultimate commerce. ISWorld. Retrieved March 2, 2004, from *www.isworld.org/ijunglas/u-commerce.htm*

Whinston, A., Barua, A., Shutter, J., Wilson, B., & Pinnell, J. (2000). Defining the Internet economy. Retrieved February 12, 2003, from *www.internet indicators.com/prod_rept.html*

KEY TERMS

Browse: To view formatted documents. For example, one looks at Web pages with a Web browser. "Browse" is often used in the same sense as "surf."

E-Commerce: Uses some form of transmission medium through which exchange of information takes place in order to conduct business.

Electronic Data Interchange: The transfer of data between different companies using networks, such as the Internet.

Electronic Fund Transfers: Any transfer of funds that is initiated through an electronic terminal, telephone, computer or magnetic tape for the purpose of ordering, instructing or authorizing a financial institution to debit or credit an account.

Internet Protocol (IP): IP specifies the format of packets, also called datagrams, and the addressing scheme. IP by itself is something like the postal system. It allows you to address a package and drop it in the system, but there is no direct link between you and the recipient.

Intranet: A network based on TCP/IP protocols belonging to an organization, usually a corporation, accessible only by the organization's members, employees or others with authorization. An intranet's Web sites look and act just like any other Web sites, but the firewall surrounding an intranet fends off unauthorized access.

Search Engine: A tool that allows a person to enter a word or phrase and then lists Web pages or items in a database that contain that phrase. The success of such a search depends on a variety of factors, including the number of Web sites that are searchable (or scope of the database), the syntax that a user enters a query in and the algorithm for determining the "relevance" of a result, which is some measure of how well a given page matches the query. A typical problem is a user retrieving too few or too many results, and having difficulty broadening or narrowing the query appropriately.

Ethernet Passive Optical Networks

Mário M. Freire
Universidade de Beira Interior, Portugal

Paulo P. Monteiro
SIEMENS S.A. and Universidade de Aveiro, Portugal

Henrique J. A. da Silva
Universidade de Coimbra, Portugal

Jose Ruela
Faculdade de Engenharia da Universidade do Porto (FEUP), Portugal

INTRODUCTION

Recently, Ethernet Passive Optical Networks (EPONs) have received a great deal of interest as a promising cost-effective solution for next-generation high-speed access networks. This is confirmed by the formation of several *fora* and working groups that contribute to their development; namely, the EPON Forum (http://www.ieeecommunities.org/epon), the Ethernet in the First Mile Alliance (http://www.efmalliance.org), and the IEEE 802.3ah working group (http://www.ieee802.org/3/efm), which is responsible for the standardization process. EPONs are a simple, inexpensive, and scalable solution for high-speed residential access, capable of delivering voice, high-speed data, and multimedia services to end users (Kramer, Mukherjee &Maislos, 2003; Kramer & Pesavento, 2002; Lorenz, Rodrigues & Freire, 2004; Pesavento, 2003; McGarry, Maier & Reisslein, 2004). An EPON combines the transport of IEEE 802.3 Ethernet frames over a low-cost and broadband point-to-multipoint passive optical fiber infrastructure connecting the Optical Line Terminal (OLT) located at the central office to Optical Network Units (ONUs), usually located at the subscriber premises. In the downstream direction, the EPON behaves as a broadcast and select shared medium, with Ethernet frames transmitted by the OLT reaching every ONU. In the upstream direction, Ethernet frames transmitted by each ONU will only reach the OLT, but an arbitration mechanism is required to avoid collisions.

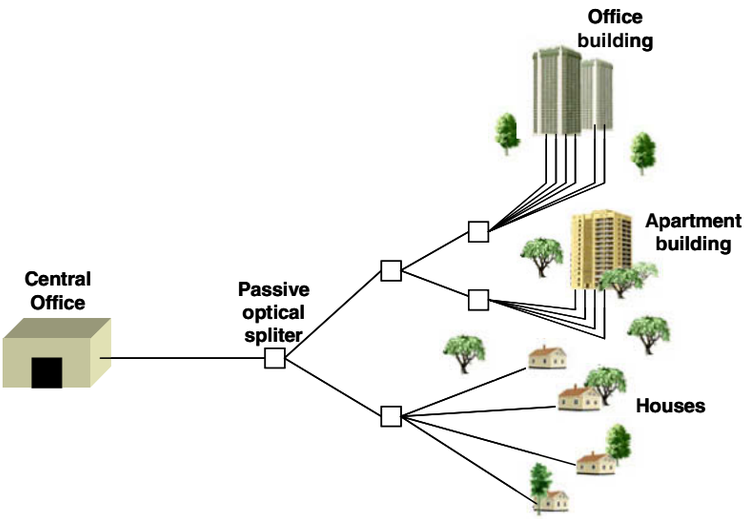
This article provides an overview of EPONs and focuses on the following issues: EPON architecture; Multi-Point Control Protocol (MPCP); quality of service (QoS); and operations, administration, and maintenance (OAM) capability of EPONs.

EPON ARCHITECTURE

EPONs, which represent the convergence of low-cost and widely used Ethernet equipment and low-cost point-to-multipoint fiber infrastructure, seem to be the best candidate for the next-generation access network (Kramer & Pesavento, 2002; Pesavento, 2003). In order to create a cost-effective shared fiber infrastructure, EPONs use passive optical splitters in the outside plant instead of active electronics, and, therefore, besides the end terminating equipment, no intermediate component in the network requires electrical power. Due to its passive nature, optical power budget is an important issue in EPON design, because it determines how many ONUs can be supported, as well as the maximum distance between the OLT and ONUs. In fact, there is a tradeoff between the number of ONUs and the distance limit of the EPON, because optical losses increase with both split count and fiber length. EPONs can be deployed to reach distances up to around 20 km with a 1:16 split ratio, which sufficiently covers the local access network (Pesavento, 2003). Figure 1 shows a possible deployment scenario for EPONs (Kramer, Banerjee, Singhal, Mukherjee, Dixit & Ye, 2004).

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

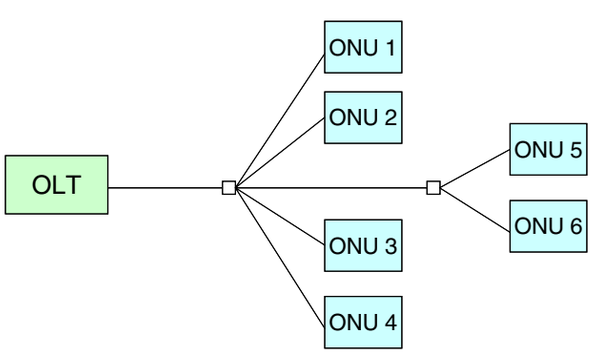
Figure 1. Schematic representation of a possible deployment scenario for EPONs



Although several topologies are possible (i.e., tree, ring, and bus) (Kramer, Mukherjee & Maislos, 2003; Kramer, Mukherjee & Pesavento, 2001; Pesavento, 2003), the most common EPON topology is a 1:N tree or a 1:N tree-and-branch network, which cascades 1:N splitters, as shown in Figure 2. The preference for this topology is due to its flexibility in adapting to a growing subscriber base and increasing bandwidth demands (Pesavento, 2003).

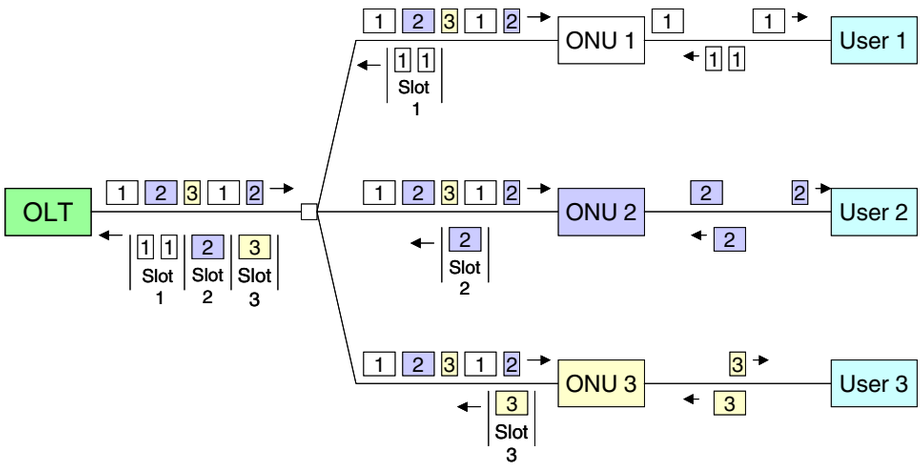
EPONs cannot be considered either a shared medium or a full-duplex point-to-point network, but a combination of both depending on the transmission direction (Pesavento, 2003). In the downstream direction, an EPON behaves as a shared medium (physical broadcast network), with Ethernet frames transmitted from the OLT to every ONU. In the upstream direction, due to the directional properties of passive couplers, which act as passive splitters for downstream, Ethernet frames from any ONU will only reach the OLT and not any other ONU. In the upstream direction, the logical behavior of an EPON is similar to a point-to-point network, but unlike in a true point-to-point network, collisions may occur among frames transmitted from different ONUs. Therefore, in the upstream direction, there is the requirement both to share the trunk fiber and to arbitrate ONU transmissions to avoid collisions by means of a Multi-Point Control Protocol (MPCP) in the Medium Access Control (MAC) layer. An overview of this protocol will be presented in the next section.

Figure 2. Schematic representation of a tree-and-branch topology for EPONs



EPONs use point-to-point emulation to meet the compliance requirements of 802.1D bridging, which provides for ONU to ONU forwarding. For this function, a 2-byte Logical Link Identifier (LLID) is used in the preamble of Ethernet frames. This 2-byte tag uses 1-bit as a mode indicator (point-to-point or broadcast mode), and the remaining 15-bits as the ONU ID. An ONU transmits frames using its own assigned LLID and receives and filters frames according to the LLID. An emulation sublayer below the Ethernet MAC demultiplexes a packet based on its LLID and strips the LLID prior to sending the frame to the MAC entity. Therefore, the LLID exists only within the EPON network. When transmitting, an LLID corresponding to the local MAC entity is added. Based on the LLID, an ONU will reject frames not intended for it. For example, a given ONU will reject

Figure 3. Illustration of frame transmission in EPONs



broadcast frames that it generates, or frames intended for other ONUs on the same PON (Pesavento, 2003).

In the downstream direction, an EPON behaves as a physical broadcast network of IEEE 802.3 Ethernet Frames, as shown in Figure 3. An Ethernet frame transmitted from the OLT is broadcast to all ONUs, which is a consequence of the physical nature of a 1:N optical splitter. At the OLT, the LLID tag is added to the preamble of each frame and extracted and filtered by each ONU in the reconciliation sublayer. Each ONU receives all frames transmitted by the OLT but extracts only its own frames; that is, those matching its LLID. Frame extraction (filtering) is based only on the LLID since the MAC of each ONU is in promiscuous mode and accepts all frames. Due to the broadcast nature of EPONs in the downstream direction, an encryption mechanism often is considered for security reasons. In the upstream direction, a multiple access control protocol is required, because the EPON operates as a physical multipoint-to-point network. Although each ONU sends frames directly to the OLT, the ONUs share the upstream trunk fiber, and simultaneous frames from ONUs might collide if the network was not properly managed. In normal operation, no collisions occur in EPONs (Pesavento, 2003).

MULTI-POINT CONTROL PROTOCOL

In order to avoid collisions in the upstream direction, EPONs use the Multi-Point Control Protocol (MPCP). MPCP is a frame-oriented protocol based on 64-byte MAC control messages that coordinate the transmission of upstream frames in order to avoid collisions. Table 1 presents the main functions performed by MPCP (Pesavento, 2003). In order to enable MPCP functions, an extension of MAC Control sublayer is needed, which is called Multipoint MAC Control sublayer.

MPCP is based on a non-cyclical frame-based Time Division Multiple Access (TDMA) scheme.

Table 1. Main functions performed by MPCP

Main Functions Performed by MPCP
• Bandwidth request and assignment • Negotiation of parameters • Managing and timing upstream transmissions from ONUs to avoid collisions • Minimization of the space between upstream slots by monitoring round trip delay • Auto-discovery and registration of ONUs

The OLT sends GATE messages to ONUs in the form of 64-byte MAC Control frames. The GATE messages contain a timestamp and granted timeslot assignments, which represent the periods in which a given ONU can transmit. The OLT allocates time slots to the ONUs. Depending on the scheduler algorithm, bandwidth allocation can be static or dynamic. It is not allowed frame fragmentation within the upstream time slot, which contains several IEEE 802.3 Ethernet frames.

For upstream operation, the ONU sends REPORT messages, which contain a timestamp for calculating round trip time (RTT) at the OLT, and a report on the status of the queues at the ONU, so that efficient dynamic bandwidth allocation (DBA) schemes can be used. The ONU is not synchronized, nor does it have knowledge of delay compensation. Moreover, for upstream transmission, the ONU transceiver receives a timely indication from MPCP to change between on and off states (Pesavento, 2003).

QUALITY OF SERVICE IN EPONs

In a multi-service network, the allocation of resources to competing users/traffic flows must provide differentiated QoS guarantees to traffic classes, while keeping efficient and fair use of shared resources. Depending on the class, guaranteed throughput or assured bounds on performance parameters, such as packet delay and jitter and packet loss ratio, may be negotiated.

In an EPON access network, sharing of the upstream and downstream channels for the communication between the OLT and a number of ONUs requires a separate analysis. In the downstream direction, the OLT is the single source of traffic (point-to-multipoint communication) and has control over the entire bandwidth of the broadcast channel; thus, resource (bandwidth) management reduces to the well-known problem of scheduling flows organized in a number of queues associated to different traffic classes. However, in the upstream direction, ONUs must share the transmission channel (multipoint-to-point communication). Besides an arbitration protocol for efficient access to the medium, it is also necessary to allocate bandwidth and schedule different classes of flows, both within each ONU and among competing ONUs, such that QoS objectives are met.

These goals may be fulfilled by means of a strategy based on MPCP and other mechanisms that take advantage of MPCP features. Since MPCP is a link layer protocol, appropriate mappings between link layer and network layer QoS parameters are required in the framework of the QoS architectural model adopted (e.g., IntServ [Braden, Clark & Shenker, 1994] or DiffServ [Blake, Black, Carlson, Davies, Wang & Weiss, 1998; Grossman, 2002]). In this article, only the lower layer mechanisms related with MPCP are discussed.

It must be stressed that MPCP is not a bandwidth allocation mechanism and does not impose or require a specific allocation algorithm. MPCP is simply a Medium Access Control (MAC) protocol based on request and grant messages (REPORT and GATE, respectively) exchanged between ONUs and the OLT. As such, it may be used to support any allocation scheme aimed at efficient and fair share of resources and provision of QoS guarantees.

The MPCP gated mechanism arbitrates the transmission from multiple nodes by allocating a transmission window (time-slot) to each ONU. Since the OLT assigns non-overlapping slots to ONUs, collisions are avoided, and, thus, efficiency can be kept high. However, this is not enough; the allocation algorithm also should avoid waste of resources (that may occur if time-slots are not fully utilized by ONUs) and support the provision of differentiated QoS guarantees to different traffic classes in a fair way.

In fact, a static allocation of fixed size slots may become highly inefficient with variable bit rate traffic, which is typical of bursty data services and many real-time applications, and with unequal loads generated by the ONUs. The lack of statistical multiplexing may lead to overflow of some slots, even under light loads, due to traffic burstiness, as well as to slot underutilization since, in this case, it is not possible to reallocate capacity assigned to and not used by an ONU. Therefore, inter-ONU scheduling based on the dynamic allocation of variable size slots to ONUs is essential both to keep the overall throughput of the system high and to fulfill QoS requirements in a flexible and scalable way.

In a recent survey, McGarry, Maier, and Reisslein (2004) proposed a useful taxonomy for classifying dynamic bandwidth allocation (DBA) algorithms for EPONs. Some only provide statistical multiplexing, while others offer QoS guarantees. The latter cat-

egory may be further subdivided into algorithms with absolute and relative QoS assurances. Some examples follow.

Kramer, Mukherjee, and Pesavento (2002) proposed an interleaved polling mechanism with adaptive cycle time (IPACT) and studied different allocation schemes. They concluded that best performance was achieved with the limited service—the OLT grants to each ONU the requested number of bytes in each polling cycle up to a predefined maximum. However, cycle times are of variable length, and, therefore, the drawback of this method is that delay jitter cannot be tightly controlled. A control theoretic extension of IPACT aimed at improving the delay performance of the algorithm has been studied by Byun, Nho, and Lim (2003).

The original IPACT scheme simply provided statistical multiplexing but did not support QoS differentiation to traffic classes. However, in a multi-service environment, each ONU has to transmit traffic belonging to different classes, and, therefore, QoS differentiation is required; this means that intra-ONU scheduling is also necessary. Incoming traffic from the users served by an ONU must be organized in separate queues based on a process that classifies and assigns packets to the corresponding traffic classes. Packets may be subject to marking, policing, and dropping, which is in conformance with a Service Level Agreement (SLA). Intra-ONU scheduling is usually based on some variant of priority queuing.

The combination of the limited service scheme and priority queuing (inter-ONU and intra-ONU scheduling, respectively) has been exploited by Kramer, Mukherjee, Dixit, Ye, and Hirth (2002) as an extension to IPACT. However, some fairness problems were identified, especially the performance degradation of some (low-priority) traffic classes when the network load decreases (a so-called light load penalty). This problem is overcome in the scheme proposed by Assi, Ye, Dixit, and Ali (2003), which combines non-strict priority scheduling with a dynamic bandwidth allocation mechanism based on but not confined to the limited service. The authors also consider the possibility of delegating into the OLT the responsibility of per-class bandwidth allocation for each ONU, since MPCP control messages can carry multiple grants. In this way, the OLT will be able to perform a more accurate allocation, based on the knowledge of per class requests sent by each ONU, at the expense of a higher complexity. This idea had been previously included in the DBA algorithm proposed by Choi and Huh (2002).

These algorithms only provide relative QoS assurances, like the two-layer bandwidth allocation algorithm proposed by Xie, Jiang, and Jiang (2004) and the dynamic credit distribution (D-CRED) algorithm described by Miyoshi, Inoue, and Yamashita (2004). Examples of DBA algorithms that offer absolute QoS assurances include Bandwidth Guaranteed Polling (Ma, Zhu & Cheng, 2003) and Deterministic Effective Bandwidth (Zhang, An, Youn, Yeo &Yang, 2003).

In spite of the progress that has been achieved in recent years, more research on this topic is still required, addressing, in particular, the optimization of scheduling algorithms combined with other QoS mechanisms, tuning of critical parameters in real operational conditions and appropriate QoS parameter mappings across protocol layers, and integration of the EPON access mechanisms in a network-wide QoS architecture aimed at the provision of end-to-end QoS guarantees.

OPERATIONS, ADMINISTRATION, AND MAINTENANCE CAPABILITY OF EPONs

OAM capability provides a network operator with the ability to monitor the network and determine failure locations and fault conditions. OAM mechanisms defined for EPONs include remote failure indication, remote loopback, and link monitoring. Remote failure indication is used to indicate that the reception path of the local device is non-operational. Remote loopback provides support for frame-level loopback and a data link layer ping. Link monitoring provides event notification with the inclusion of diagnostic data and polling of variables in the IEEE 802.3 Management Information Base. A special type of Ethernet frames called OAM Protocol Data Units, which are slow protocol frames, are used to monitor, test, and troubleshoot links. The OAM protocol also is able to negotiate the set of OAM functions that are operable on a given link interconnecting Ethernet devices (Pesavento, 2003).

CONCLUSION

EPONs have been proposed as a cost-effective solution for next-generation high-speed access networks. An overview of major issues in EPONs has been presented. The architecture and principle of operation of EPONs were briefly described. The Multi-Point Control Protocol used to eliminate collisions in the upstream direction was briefly presented. Quality of service, a major issue for multimedia services in EPONs, was also addressed. The operations, administration, and maintenance capability of EPONs was also briefly discussed.

REFERENCES

Assi, C.M., Ye, Y., Dixit, S., & Ali, M.A. (2003). Dynamic bandwidth allocation for quality-of-service over ethernet PONs. *IEEE Journal on Selected Areas in Communications, 21*(9), 1467-1477.

Blake, S., et al. (1998). An architecture for differentiated services. *Internet Engineering Task Force*, RFC 2475.

Braden, R., Clark, D., & Shenker, S. (1994). Integrated services in the Internet architecture: An overview. *Internet Engineering Task Force*, RFC 1633.

Byun, H.-J., Nho, J.-M., & Lim, J.-T. (2003). Dynamic bandwidth allocation algorithm in ethernet passive optical networks. *IEE Electronics Letters, 39*(13), 1001-1002.

Choi, S.-I., & Huh, J.-D. (2002). Dynamic bandwidth allocation algorithm for multimedia services over ethernet PONs. *ETRI Journal, 24*(6), 465-468.

Grossman, D. (2002). New terminology and clarifications for Diffserv. *Internet Engineering Task Force*, RFC 3260.

Kramer, G., et al. (2004). Fair queuing with service envelopes (FQSE): A cousin-fair hierarchical scheduler for ethernet PON. *Proceedings of the Optical Fiber Communications Conference (OFC 2004),* Los Angeles.

Kramer, G., & Pesavento, G. (2002). Ethernet passive optical network (EPON): Building a next-generation optical access network. *IEEE Communications Magazine, 40*(2), 68-73.

Kramer, G., Mukherjee, B., & Maislos, A. (2003). Ethernet passive optical networks. In S. Dixit (Ed.), *Multiprotocol over DWDM: Building the next generation optical Internet*. Hoboken, NJ: John Wiley & Sons.

Kramer, G., Mukherjee, B., & Pesavento, G. (2001). Ethernet PON (ePON): Design and analysis of an optical access network. *Photonic Network Communications, 3*(3), 307-319.

Kramer, G., Mukherjee, B., & Pesavento, G. (2002). IPACT: A dynamic protocol for an ethernet PON (EPON). *IEEE Communications Magazine, 40*(2), 74-80.

Kramer, G., Mukherjee, B., Dixit, S., Ye, Y., & Hirth, R. (2002). Supporting differentiated classes of service in ethernet passive optical networks. *Journal of Optical Networking, 1*(8-9), 280-298.

Lorenz, P., Rodrigues J.J.P.C., & Freire, M.M. (2004). Fiber-optic networks. In R. Driggers (Ed.), *Encyclopedia of optical engineering*. New York: Marcel Dekker.

Ma, M., Zhu, Y., & Cheng, T.H. (2003). A bandwidth guaranteed polling MAC protocol for ethernet passive optical networks. *Proceedings of IEEE INFOCOM 2003,* (Vol. 1, pp. 22-31).

McGarry, M.P., Maier, M., & Reisslein, M. (2004). Ethernet PONs: A survey of dynamic bandwidth allocation (DBA) algorithms. *IEEE Optical Communications, 2*(3), S8-S15.

Miyoshi, H., Inoue, T., & Yamashita, K. (2004). QoS-aware dynamic bandwidth allocation scheme in gigabit-ethernet passive optical networks. *Proceedings of IEEE International Conference on Communications*, (Vol. 1, pp. 90-94). Paris, France.

Pesavento, G. (2003). Ethernet passive optical network (EPON) architecture for broadband access. *Optical Networks Magazine, 4*(1), 107-113.

Xie, J., Jiang, S., & Jiang, Y. (2004). A dynamic bandwidth allocation scheme for differentiated services in EPONs. *IEEE Optical Communications, 2*(3), S32-S39.

Zhang, L., An, E.-S., Youn, C.-H., Yeo, H.-G., & Yang, S. (2003). Dual DEB-GPS scheduler for delay-constraint applications in ethernet passive optical networks. *IEICE Transactions on Communications, E86-B*(5), 1575-1584.

KEY TERMS

DBA: Dynamic Bandwidth Allocation. DBA algorithms can be used with the MPCP arbitration mechanism to determine the collision-free upstream transmission schedule of ONUs and generate GATE messages accordingly.

Ethernet Frame: It consists of a standardized set of bits, organized into several fields, used to carry data over an Ethernet system. Those fields include the preamble, a start frame delimiter, address fields, a length field, a variable size data field that carries from 46 to 1,500 bytes of data, and an error checking field.

LLID: Logical Link Identifier. LLID is a 2-byte tag in the preamble of an Ethernet frame. This 2-byte tag uses 1-bit as a mode indicator (point-to-point or broadcast mode) and the remaining 15-bits as the ONU ID.

MPCP: Multi-Point Control Protocol. Medium access control protocol used in EPONs to avoid collisions in the upstream direction.

OLT: Optical Line Terminal. An OLT is located at the central office and is responsible for the transmission of Ethernet frames to ONUs.

ONU: Optical Network Unit. An ONU is usually located at the subscriber premises or in a telecom closet and is responsible for the transmission of Ethernet frames to OLT.

PON: Passive Optical Network. A PON is a network based on optical fiber in which all active components and devices between the central office and the customer premises are eliminated.

Evolution of GSM Network Technology

Phillip Olla
Brunel University, UK

INTRODUCTION

The explosive growth of Global System for Mobile (GSM) Communication services over the last two decades has changed mobile communications from a niche market to a fundamental constituent of the global telecommunication markets. GSM is a digital wireless technology standard based on the notion that users want to communicate wirelessly without limitations created by network or national borders. In a short period of time, GSM has become a global phenomenon. The explanation for its success is the cooperation and coordination of technical and operational evolution that has created a virtuous circle of growth built on three principles: interoperability based on open platforms, roaming, and economies of scale (GSM Association, 2004a). GSM standards are now adopted by more than 200 countries and territories. It has become the main global standard for mobile communications; 80% of all new mobile customers are on GSM networks. GSM has motivated wireless adoption to the extent that mobile phones now globally outnumber fixed-line telephones. In February 2004, more than 1 billion people, almost one in six of the world's population, were using GSM mobile phones.

Some developed European nations such as the United Kingdom, Norway, Finland, and Spain have penetration levels of between 80 to 90% with other European nations not far behind. However, there are some countries such as Hong Kong and Italy that have a 100% penetration level. The importance of the mobile telecommunication industry is now apparent: A recent study commissioned by a UK mobile operator establishes that the United Kingdom's mobile-phone sector now contributes as much to the UK gross domestic product as the total oil- and gas-extraction industry (MMO2, 2004).

Technical developments, competition, and deregulation have contributed to a strong growth in the adoption of mobile phones in the third world. In Africa, recent research has shown that mobile telephony has been extremely important in providing an African telecommunications infrastructure. The number of mobile phone users on the African continent has increased by over 1,000% between 1998 and 2003 to reach a total of 51.8 million. Mobile-user numbers have exceeded those of fixed line, which stood at 25.1 million at the end of 2003. The factors for success in this region include demand, sector reform, the licensing of new competition, and the emergence of important strategic investors (ITU, 2004). Another region experiencing rapid growth is India; it is one of the fastest growing markets, with its subscriber base doubling in 2003. It is anticipated that India will have 100 million GSM subscribers by 2007 and 2008 compared to 26 million subscribers as of March 2004 (3G Portal, 2004). Most Latin American operators have chosen GSM over the North American code-division multiple-access (CDMA) standards, and GSM growth in North America is higher than CDMA.

This article describes the evolution of the telecommunication networks from the first-generation networks of the '80s to the revolutionary fourth-generation networks.

FOCUS: EVOLUTION OF GSM NETWORKS

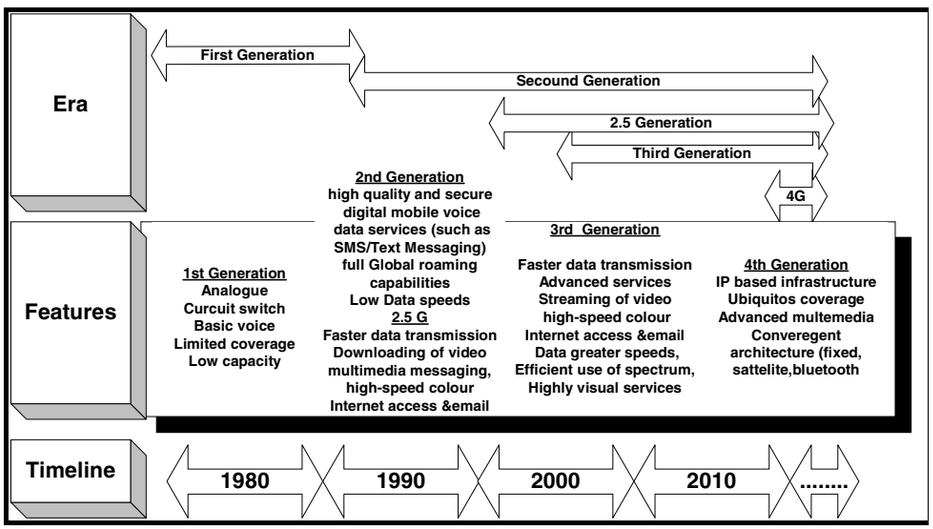
Mobile communications can be divided into three distinct eras identified by an increase in functionality and bandwidth, as illustrated in Figure 1. These eras relate to the implementation of technological advancements in the field. The industry is currently on the verge of implementing the third technological era and at the beginning of defining the next step for the fourth era.

First-Generation Networks

The first-generation (1G) cellular systems were the simplest communication networks deployed in the 1980s. The first-generation networks were based on

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Mobile telecommunication eras



analogue-frequency-modulation transmission technology. Challenges faced by the operators included inconsistency, frequent loss of signals, and low bandwidth. The 1G network was also expensive to run due to a limited customer base.

Second-Generation Networks

The second-generation (2G) cellular systems were the first to apply digital transmission technologies for voice and data communication. The data transfer rate was in the region of 10s of Kbps. Other examples of technologies in 2G systems include frequency-division multiple access (FDMA), time-division multiple access (TDMA), and code-division multiple access.

The second-generation networks deliver high-quality and secure mobile voice, and basic data services such as fax and text messaging along with full roaming capabilities across the world.

To address the poor data transmission rates of the 2G network, developments were made to upgrade 2G networks without replacing the networks. These technological enhancements were called 2.5G technologies and include networks such as General Packet Radio Service (GPRS). GPRS-enabled networks deliver features such as always-on, higher capacity, Internet-based content and packet-based data services enabling services such as colour Internet browsing, e-mail on the move, visual communications, multimedia messages, and location-based services. Another complementary 2.5G service is Enhanced Data Rates for GSM Evolution (EDGE). This network upgrade offers similar capabilities as those of the GPRS network. Another 2.5G network enhancement of data services is high-speed circuit-switched data (HSCSD). This allows access to nonvoice services 3 times faster than conventional networks, which means subscribers are able to send and receive data from their portable computers at speeds of up to 28.8 Kbps; this is currently being upgraded in many networks to 43.2 Kbps. The HSCSD solution enables higher rates by using multiple channels, allowing subscribers to enjoy faster rates for their Internet, e-mail, calendar, and file-transfer services. HSCSD is now available to more than 100 million customers across 27 countries around the world in Europe, Asia Pacific, South Africa, and Israel (GSM, 2002)

Current Trend: Third-Generation Networks

The most promising period is the advent of third-generation (3G) networks. These networks are also referred to as the universal mobile telecommunications systems (UMTSs). The global standardization effort undertaken by the ITU is called IMT-2000. The aim of the group was to evolve today's circuit-

switched core network to support new spectrum allocations and higher bandwidth capability. Over 85% of the world's network operators have chosen 3G as the underlying technology platform to deliver their third-generation services (GSM, 2004b).

The implementation of the third generation of mobile systems has experienced delays in the launch of services. There are various reasons for the delayed launch, ranging from device limitations, application- and network-related technical problems, and lack of demand. A significant factor in the delayed launch that is frequently discussed in the telecommunication literature (Klemperer, 2002; Maitland, Bauer, & Westerveld, 2002; Melody, 2000) is the extortionate fees paid for the 3G-spectrum license in Europe during the auction process. Most technical problems along with device shortage have been overcome, but there are still financial challenges to be addressed caused by the high start-up costs and the lack of a subscriber base due to the market saturation in many of the countries launching 3G.

In 2002, industry experts revealed lower-than-expected 3G forecasts. The continued economic downturn prompted renewed concerns about the near-term commercial viability of mobile data services, including 3G. The UMTS forum reexamined the worldwide market demand for 3G services due to the effects of September 11 and the global telecommunication slump, and produced an updated report (UMTS, 2003).

The reexamination highlighted the fact that due to the current negative market conditions, the short-term revenue generated by 3G services will be reduced by 17% through 2004: a total reduction of $10 billion. However, over the long term, services enabled by 3G technology still represent a substantial market opportunity of $320 billion by 2010, $233 billion of which will be generated by new 3G services (Qiu & Zhang, 2002).

Future Trends: Fourth-Generation Mobile Networks

The fourth-generation (4G) systems are expected around 2010 to 2015. They will be capable of combining mobility with multimedia-rich content, high bit rates, and Internet-protocol (IP) transport.

The benefits of the fourth-generation approach are described by Inforcom Research (2002) and Qiu et al. (2002) as voice-data integration, support for mobile and fixed networking, and enhanced services through the use of simple networks with intelligent terminal devices. The fourth-generation networks are expected to offer a flexible method of payment for network connectivity that will support a large number of network operators in a highly competitive environment.

Over the last decade, the Internet has been dominated by non-real-time, person-to-machine communications. According to a UMTS report (2002b), the current developments in progress will incorporate real-time, person-to-person communications, including high-quality voice and video telecommunications along with the extensive use of machine-to-machine interactions to simplify and enhance the user experience.

Currently, the Internet is used solely to interconnect computer networks; IP compatibility is being added to many types of devices such as set-top boxes, automotive systems, and home electronics. The large-scale deployment of IP-based networks will reduce the acquisition costs of the associated devices. The future vision is to integrate mobile voice communications and Internet technologies, bringing the control and multiplicity of Internet-applications services to mobile users.

The creation and deployment of IP-based multimedia services (IMSs) allows person-to-person real-time services, such as voice over the 3G packet-switched domain (UMTS, 2002a). IMS enables IP interoperability for real-time services between fixed and mobile networks, solving current problems of seamless, converged voice-data services. Service transparency and integration are key features for accelerating end-user adoption. Two important features of IMS are IP-based transport for both real-time and non-real-time services, and a multimedia call-model based on the session-initiation protocol (SIP). The deployment of an IP-based infrastructure will encourage the development of voice-over-IP (VoIP) services.

The current implementation of the Internet protocol, Version 4 (IPv4), is being upgraded due to the constraints of providing new functionality for modern devices. The pool of Internet addresses is also being depleted. The new version, called IP, Version 6 (IPv6), resolves IPv4 design issues and is primed to take the Internet to the next generation. Internet protocol, Version 6, is now included as part of IP

support in many products including the major computer operating systems.

CONCLUSION

In just over two decades, mobile network technologies have evolved from simple 1G networks to today's 3G networks, which are capable of high-speed data transmission allowing innovative applications and services. The evolution of the communication networks is fueling the development of the mobile Internet and creating new types of devices. In the future, 4G networks will supersede 3G.

The fourth-generation technology supports broadly similar goals to the third-generation effort, but starts with the assumption that future networks will be entirely packet-switched using protocols evolved from those in use in today's Internet. Today's Internet telephony systems are the foundation for the applications that will be used in the future to deliver innovative telephony services.

REFERENCES

3G Portal. (2004). *India: Driving GSM to the next billion subscribers.* Retrieved from *http://www.the3gportal.com/3gpnews/archives/007143.html#007143*

GSM Association. (2002, March 12). *High-speed data communication now available to over 100 million GSM users in 27 countries worldwide* [Press release].

GSM Association. (2004a). *GSM Association brochure.* Retrieved from *http://www.gsmworld.com/news/newsletter.shtml*

GSM Association. (2004b). *GSM information.* Retrieved from *http://www.gsmworld.com/index.shtml*

Inforcom Research. (2002). The dawn of 3.5 and 4G next generation systems. *Gateway to N+1 Generation Networks, 1*(4). Retrieved from *http://www.icr.co.jp/nG/src/0104_contents.pdf*

ITU. (2004). Africa: The world's fastest growing mobile market. Does mobile technology hold the key to widening access to ICTs in Africa? *African Telecommunication Indicators 2004*.

Klemperer, P. (2002). How (not) to run auctions: The European 3G telecom auctions. *European Economic Review, 46*(4-5), 829-845.

Maitland, C. F., Bauer, J. M., & Westerveld, R. (2002). The European market for mobile data. *Telecommunications Policy, 26*(9-10), 485-504.

Melody, W. H. (2000). Telecom development. *Telecommunications Policy, 24*(8-9), 635-638.

MMO2. (2004). Mobile communications a vital contributor to global. Retrieved from *http://www.gsmworld.com/index.shtml*

Qiu, R. C., W. Z., & Zhang, Y. Q. (2002). *Third-generation and beyond (3.5G) wireless networks and its applications* IEEE International Symposium on Circuits and Systems (ISCS), Scottsdale, AZ.

UMTS. (2002a). *IMS service vision for 3G markets* (Forum Rep. 20). Retrieved from *http://www.umts-forum.org/servlet/dycon/ztumts/umts/Live/en/umts/Resources_Reports_index:*

UMTS. (2002b). *Support of third generation services using UMTS in a converging network environment* (Forum Rep. 14). Retrieved from *http://www.umts-forum.org/servlet/dycon/ztumts/umts/Live/en/umts/Resources_Reports_index:*

UMTS. (2003). *The UMTS 3G market forecasts: Post September 11, 2001* (Forum Rep. 18). Retrieved from *http://www.umts-forum.org/servlet/dycon/ztumts/umts/Live/en/umts/Resources_Reports_index:*

KEY TERMS

Bandwidth: In networks, bandwidth is often used as a synonym for data transfer rate: the amount of data that can be carried from one point to another in a given time period (usually a second). This kind of bandwidth is usually expressed in bits (of data) per second (bps).

Circuit Switched: Circuit switched is a type of network in which a physical path is obtained for and dedicated to a single connection between two end-

points in the network for the duration of the connection. Ordinary voice phone service is circuit switched. The telephone company reserves a specific physical path to the number you are calling for the duration of your call. During that time, no one else can use the physical lines involved.

EDGE: Enhanced Data Rates for GSM Evolution, a faster version of the GSM wireless service, is designed to deliver data at rates up to 384 Kbps and enable the delivery of multimedia and other broadband applications to mobile phone and computer users.

GPRS: General Packet Radio Service (GPRS) is a packet-based wireless communication service that promises data rates from 56 up to 114 Kbps, and continuous connection to the Internet for mobile phone and computer users. The higher data rates will allow users to take part in videoconferences and interact with multimedia Web sites and similar applications using mobile handheld devices as well as notebook computers.

GSM: Global System for Mobile Communication is a digital mobile telephone system that is widely used in Europe and other parts of the world. GSM uses a variation of time-division multiple access (TDMA) and is the most widely used of the three digital wireless telephone technologies (TDMA, GSM, and CDMA). GSM digitizes and compresses data, then sends it down a channel with two other streams of user data, each in its own time slot. It operates at either the 900-MHz or 1,800-MHz frequency band.

Kbps: Kbps (or Kbits) stands for kilobits per second (thousands of bits per second) and is a measure of bandwidth (the amount of data that can flow in a given time) on a data-transmission medium. Higher bandwidths are more conveniently expressed in megabits per second (Mbps, or millions of bits per second) and in gigabits per second (Gbps, or billions of bits per second).

Mobile IPv6: MIPv6 is a protocol developed as a subset of the Internet protocol, Version 6 (IPv6), to support mobile connections. MIPv6 is an update of the IETF (Internet Engineering Task Force) mobile IP standard designed to authenticate mobile devices using IPv6 addresses.

Packet Switched: Packet switched describes the type of network in which relatively small units of data called packets are routed through a network based on the destination address contained within each packet. Breaking communication down into packets allows the same data path to be shared among many users in the network.

UMTS: Universal Mobile Telecommunications Service is a third-generation (3G) broadband, packet-based transmission of text, digitized voice, video, and multimedia at data rates up to 2 Mbps that offers a consistent set of services to mobile computer and phone users no matter where they are located in the world.

Evolution of Mobile Commerce Applications

George K. Lalopoulos
Hellenic Telecommunications Organization S.A. (OTE), Greece

Ioannis P. Chochliouros
Hellenic Telecommunications Organization S.A. (OTE), Greece

Anastasia S. Spiliopoulou-Chochliourou
Hellenic Telecommunications Organization S.A. (OTE), Greece

INTRODUCTION

The tremendous growth in mobile communications has affected our lives significantly. The mobile phone is now pervasive and used in virtually every sector of human activity—private, business, and government. Its usage is not restricted to making basic phone calls; instead, digital content, products, and services are offered. Among them, mobile commerce (m-commerce) holds a very important and promising position.

M-commerce can be defined as: using mobile technology to access the Internet through a wireless device such as a cell phone or a PDA (Personal Digital Assistant), in order to sell or buy items (products or services), conduct a transaction, and perform supply-chain or demand-chain functions (Adams, 2001).

Within the context of the present study, we shall examine widespread used and emerging m-commerce services, from early ones (i.e., SMS [Short Message Service]) to innovative (i.e., mobile banking and specific products offered by known suppliers). We shall also investigate some important factors for the development of m-commerce, as well as some existing risks. Particular emphasis is given to the issue of collaboration among the key-players for developing standardization, interoperability, and security, and for obtaining market penetration.

M-COMMERCE SERVICES AND COMMERCIAL PRODUCTS

M-commerce products and services involve a range of main players, including Telcos (telecommunications service providers), mobile operators, mobile handset manufacturers, financial institutions, suppliers, payment service providers, and customers. Each party has its own interests (e.g., Telcos and mobile operators are interested not only in selling network airtime, but also in becoming value-added services providers offering additional functionality; banks consider the adaptation of their financial services to mobile distribution channels). However, successful cooperation of the involved parties is the key to the development of m-commerce.

Today's most profitable m-commerce applications concern entertainment (e.g., SMS, EMS [Enhanced Message Service], MMS [Multimedia Message Service], ring tones, games, wallpaper, voting, gambling, etc.). However, new interactive applications such as mobile shopping, reservations and bookings, ticket purchases via mobile phones (for train and bus travel, cinemas and theaters, car parking, etc.), m-cash micro purchases (for vending machines, tollbooths, etc.), mobile generation, assignment and tracking of orders, mobile banking, brokering, insurance, and business applications (e.g., accessing corporate data) have emerged and are expected to evolve and achieve significant market penetration in the future. In addition, future m-commerce users are likely to view certain goods and services not only as m-commerce products, but also in terms of situations such as being lost or having a car break down, where they will be willing to pay more for specific services (e.g., location awareness, etc.).

Mobile banking (m-banking) is the implementation of banking and trading transactions using an Internet-enabled wireless device (e.g., mobile phones, PDAs, handheld computers, etc.) without physical presentation at a bank branch. It includes services such as balance inquiry, bill payment, transfer of funds,

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

statement request, and so forth. However, there are some problems regarding future development and evolution of mobile banking services. Many consumers consider those services difficult to use and are not convinced about their safety, while financial institutions are probably waiting for a payoff from their earlier efforts to get people to bank using their personal computers and Internet connections (Charny, 2001). As a consequence, the growth of mobile banking has been relatively slow since the launch of the first m-banking products by European players in 1999 and 2000. Currently, the main objective of mobile banking is to be an additional channel with a marginal role in a broader multi-channel strategy. Nevertheless, these strategic purposes are expected to change with the development of new applications of the wireless communication market, especially in the financial sector.

Now we will examine some characteristic m-commerce products. Japan's NTT DoCoMo was the first mobile telephone service provider to offer m-commerce services by launching the i-mode service in 1999 (NTT DoCoMo, 2004, Ryan, 2000). Key i-mode features include always-on packet connections, NTT's billing users for microcharges on behalf of content providers, and user's open access to independent content sites.

T-Mobile has developed a suite of applications called Mobile Wallet and Ticketing in the City Guide (T-Mobile, 2003). The first is a mobile payment system designed for secure and comfortable shopping. T-Mobile customers in Germany already use this system via WAP (Wireless Application Protocol). The highlight of the service is that customers do not have to provide any sensitive data like payment or credit card information when they make mobile purchases. Instead, after logging-in using personal data such as name, address, and credit card or bank details, they receive a personal identification number (PIN). By entering this PIN, a user can make a purchase from participating retailers.

With the Ticketing in the City Guide application, T-Mobile demonstrates a special future mobile commerce scenario. Here, entrance tickets for events such as concerts or sporting events can be ordered using a UMTS (Universal Mobile Telecommunications System) handset and paid for via Mobile Wallet. The tickets are sent to the mobile telephone by SMS in the form of barcodes. The barcodes can be read using a scanner at the venue of the event and checked to confirm their validity; subsequently, a paper ticket can be printed using a connected printer.

Nokia offers mobile commerce solutions such as the Nokia Payment Solution and the Wallet applications (Nokia Press Releases, 2001). The first one networks consumers, merchants, financial institutions, content/service providers, and various clearing channels in order to enable the exchange of funds among these parties and to allow users to make online payments for digital content, goods, and services via the Internet, WAP, or SMS. It collects, manages, and clears payments initiated from mobile phones and other Web-enabled terminals through various payment methods like credit and debit cards, operator's pre-paid or post-paid systems, and a virtual purse, which is an integrated pre-paid account of Nokia's Payment Solution that can be used with specific applications (e.g., mobile games). The solution enables remote payments from mobile terminals (e.g., electronic bill payment and shopping, mobile games, ticketing, auctioning, music downloading, etc.) and local payments (e.g., vending machines, parking fees, etc.).

Wallet is a password-protected area in the phone where users can store personal information such as payment card details, user names, and passwords, and easily retrieve it to automatically fill in required fields while browsing on a mobile site.

FACTORS AND RISKS

The development of advanced m-commerce applications, in combination with the evolution of key infrastructure components such as always-on high-speed wireless data networks (e.g., 2.5G, 3G, etc.) and mobile phones with multi-functionality (e.g., built-in-camera, music player, etc.) is stimulating the growth of m-commerce. Other key drivers of m-commerce are ease-of-use, convenience, and anytime-anywhere availability. On the other hand, a customer's fear of fraud is a major barrier. The nature of m-commerce requires a degree of trust and cooperation among member nodes in networks that can be exploited by malicious entities to deny service, as well as collect confidential information and disseminate false information. Another obvious risk is loss or theft of mobile devices. Security, therefore, is absolutely necessary

for the spreading of m-commerce transactions with two main enablers:

- A payment authentication to verify that the authorized user is making the transaction; and
- Wireless payment-processing systems that make it possible to use wireless phones as point-of-sale terminals.

These elements of security are fundamental in order to gain consumer trust.

Mobile phones can implement payment authentication through different solutions: single chip (authentication functionality and communication functionality integrated in one chip—SIM [Subscriber Identification Module]); dual chip (separate chips for authentication and communication); and dual slot (authentication function is built in a carrier card separate from the mobile device, and an external or internal card reader intermediates the communication of the card and the mobile device) (Zika, 2004).

Furthermore, several industry standards have been developed: WAP, WTLS (Wireless Transport Layer Security), WIM (Wireless Identity Module), and so forth. In particular, as far as authentication is concerned, many security companies have increased their development efforts in wireless security solutions such as Public Key Infrastructure (PKI), security software (Mobile PKI), digital signatures, digital certificates, and smart-card technology (Centeno, 2002). PKI works the same way in a wireless environment as it does in the wireline world, with more efficient usage of available resources (especially bandwidth and processing power) due to existing limitations of wireless technology. Smart-card technology allows network administrators to identify users positively and confirm a user's network access and privileges. Today, mobile consumers are using smart cards for a variety of activities ranging from buying groceries to purchasing movie tickets. These cards have made it easier for consumers to store information securely, and they are now being used in mobile banking, health care, telecommuting, and corporate network security. An example of a security mechanism is the Mobile 3-D Secure Specification developed by Visa International (Cellular Online, Visa Mobile, 2004; Visa International, 2003).

New advanced mobile devices have tracking abilities that can be used to deliver location-specific targeted advertisements or advanced services (e.g., directions for traveling, information about location of the nearest store, etc.). This additional convenience, however, has its risks due to its intrusive nature, since tracking technology may be seen as an invasion of privacy and a hindrance to an individual's ability to move freely (the "Big Brother" syndrome).

The existence of many different solutions for m-commerce leads to a need for standardization, which can result from market-based competition, voluntary cooperation, and coercive regulation.

Voluntary Cooperation

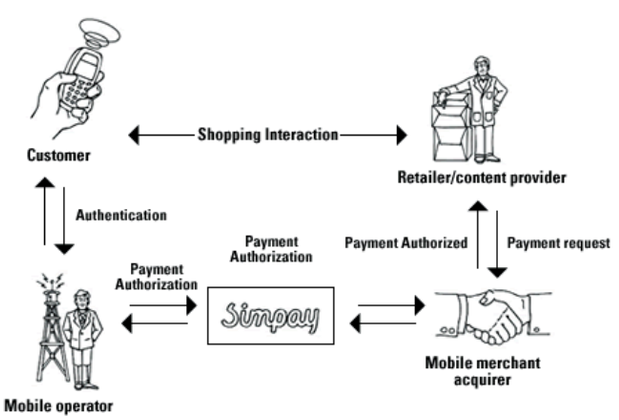
Some significant forums for the development of m-commerce are the following:

- **Mobile Payment Forum (http://www.mobilepaymentforum.org/):** A global, cross-industry organization aiming to develop a framework for secure, standardized, and authenticated mobile payment that encompasses remote and proximity transactions, as well as micro-payments. It also is taking a comprehensive approach to the mobile payments process and creating standards and best practice for every phase of a payment transaction, including the setup and configuration of the mobile payment devices, payment initiation, authentication, and completion of a transaction. Members include American Express, Master Card, Visa, Japan Card Bureau, Nokia, TIM, and so forth.
- **MeT—Mobile Electronic Transaction (http://www.mobiletransaction.org/):** It was founded to establish a common technology framework for secure mobile transactions, ensuring a consistent user experience independent of device, service, and networks, and building on existing industry security standards such as evolving WAP, WTLS, and local connectivity standards such as Bluetooth. Members include Ericsson, Motorola, Nokia, Siemens, Sony, Wells Fargo Bank, Verisign, Telia, and so forth.
- **Mobey Forum (http://www.mobeyforum.org/):** A financial industry-driven forum, whose mission is to encourage the use of mobile technology in financial services. Activities include consolidation of business and

security requirements, evaluation of potential business models, technical solutions, and recommendations to key-players in order to speed up the implementation of solutions. Members include ABN AMRO Bank, Deutsche Bank, Ericsson, Nokia, Siemens, Accenture, NCR, and so forth.

- **Open Mobile Alliance (OMA) (http://www.openmobilealliance.org/):** The mission of OMA is to deliver high-quality, open technical specifications based upon market requirements that drive modularity, extensibility, and consistency among enablers, in order to guide industry implementation efforts and provide interoperability across different devices, geographies, service providers, operators, and networks. Members include Bell Canada, British Telecommunications, Cisco Systems, NTT DoCoMo, Orange, Lucent Technologies, Microsoft Corporation, Nokia, and so forth.
- **Simpay (http://www.simpay.com/):** In order to facilitate mobile payments and deal with the lack of a single technical standard open to all carriers, four incumbent carriers (Orange, Telefonica Moviles, T-Mobile, and Vodafone) founded a consortium called Simpay (formerly known as Mobile Services Payment Association [MPSA]). Simpay was created to drive m-commerce through the development of an open and interoperable mobile payment solution, providing clearance and settlement services and a payment scheme that allow customers to make purchases through mobile-operator-managed accounts (see Figure 1).

Figure 1. Simpay's mobile payment solution



The mobile merchant acquirer (MA), after signing an agreement with Simpay, aggregates merchants (e-commerce sites that sell goods or services to the customer [in Figure1, retailers/content providers]) by signing them up and integrating them with the scheme. Any industry player (i.e., mobile operators, financial institutions, portals, etc.) can become an MA, provided that they have passed the certification and agree on the terms and conditions contractually defined by Simpay.

Membership in Simpay includes mobile operators and other issuers of SIM cards such as service providers and Mobile Virtual Network Operators (MVNOs).

When the customer clicks the option to pay with Simpay, the mobile operator provides details of the transaction to the customer's mobile phone screen. The customer clicks to send confirmation. Simpay then routes the payment details (the payment request and the authorization) between the mobile operator (a Simpay member) and the merchant acquirer who, in turn, interacts with the merchant. Purchases will be charged to the customer's mobile phone bill or to a pre-paid account with the customer's particular operator.

The technical launch for Simpay was expected at the end of 2004 and the commercial one early in 2005 (Cellular Online, Simpay Mobile, 2004). At launch, Simpay would focus on micropayments of under 10 euros for digital content (e.g., java games, ringtones, logos, video clips, and MP3 files). Higher-priced items such as flights and cinema tickets with billing to credit or debit cards will follow.

- **Wireless Advertising Association (http://www.waaglobal.org/):** An independent body that evaluates and recommends standards for mobile marketing and advertising, documents advertising effectiveness, and educates the industry on effective and responsible methods. Members include AT&T Wireless, Terra Lycos, Nokia, AOL Mobile, and so forth.

Regulation

Directives from the authorities can boost consumer trust in m-commerce. This is the case in Japan, where regulators have set up standards for operators who wish to offer m-payment facilities to their users. The

system also requires companies who allow for mobile payments to be registered with government regulations, so that consumers know they can get a refund if a service is not delivered as promised (Clark, 2003).

EU Directives

The European Commission has proposed some directives in an effort to harmonize regulatory practices of member countries. In September 2000, two directives on e-money were adopted: the ELMI Directive (Directive 46/EC, 2000) of the European Parliament and the Council of 18 September 2000 on the taking up, pursuit of, and prudential supervision of the business of electronic money institutions; and Directive 28/EC, 2000 of the European Parliament and the Council of 18 September 2000, amending Directive 12/EC, 2000 relating to the taking up and pursuit of the business of credit institutions.

The e-money Directives introduced a set of harmonized prudential rules that should be adopted by national regulators. By implementing these requirements, the national regulators would be allowed to authorize and supervise e-money issuers that could enter the whole market of the EU without the necessity of authorization in other countries (Zika, 2004). This strategy, however, might create some problems due to the wide disparity in implementation from country to country in the EU (e.g., e-money issuers in Italy have strict regulatory demands compared to the relatively laissez-faire attitude toward regulation of mobile transactions in Finland). Consequently, some EU members' mobile payments and related content services infrastructure could develop much more quickly than others, based solely on a country's legislative approach to implementation of supposedly standard Europe-wide legislation. Therefore, a balanced approach is needed in order to facilitate competition and to develop mobile business throughout Europe, toward smoothing the existing differences between different countries in the EU (EU Information Society Portal, 2003).

Moreover, under the umbrella of the e-Europe 2005 Action Plan, which is part of the strategy set out at the Lisbon European Council to modernize the European economy and to build a knowledge-based economy in Europe, a blueprint on mobile payments is under development (working document). This blueprint aims at providing a broadly supported approach that could give new momentum to industry-led initiatives and accelerate the large-scale deployment of sustainable mobile payment services, including pre-paid, post-paid, and online services, as well as payments at the point-of-sale (e-Europe Smart Card, 2003).

The EU Blueprint formally supports two objectives of the Action Plan eEurope 2005, which sets the scene for a coordinated European policy approach on information society issues:

- Interoperability
- Reduce barriers to broadband deployment (including 3G communications)

Issues like security and risk management, technical infrastructure, regulation and oversight of payment services provision, stimulation and protection of investments, and independence of mobile services providers from mobile networks are examined within the scope of the blueprint, which is expected to be endorsed by the main stakeholders (i.e., critical mass of market actors in both the financial and telecommunications sectors, as well as the relevant public authorities) by the end of 2005.

Regulation in the U.S.

The U.S. approach, in contrast to that of the EU, is based on a more relaxed view of e-money. From the beginning, the Federal Reserve (Fed) pointed out that early regulation might suppress innovation. This does not imply, however, that the regulatory interventions in the U.S. are minimal compared to the EU. In fact, besides the great number of regulatory and supervisory agencies applying a broad range of very confined rules, there also are many regulators at the state and federal level. Among them, the Uniform Money Services Act (UMSA) aims at creating a uniform legal framework in order to give non-banks the opportunity to comply with the various state laws when conducting business on a nationwide level. UMSA covers a wide range of financial (payment) services, not just e-money activities (Zika, 2004).

CONCLUSION

Mobile commerce (m-commerce) is seen as a means to facilitate certain human activities (i.e., entertainment, messaging, advertising, marketing, shopping, information acquisition, ticket purchasing, mobile banking, etc.), while offering new revenue opportunities to involved parties in the related supply chain (i.e., mobile operators, merchants/retailers, service providers, mobile handset manufacturers, financial institutions, etc.).

However, there are some barriers preventing m-commerce from taking off. They include lack of user trust in m-commerce technology, doubts about m-commerce security, and lack of widely accepted standards. As a consequence, the main income source for today's m-commerce services is the entertainment sector with low-price applications such as ringtones, wallpapers, games, lottery, horoscopes, and so forth.

With the advent of high-speed wireless networks (e.g., 2.5G, 3G, etc.) and the development of advanced applications such as mobile shopping, mobile ticketing, mobile banking, and so forth, m-commerce is expected to take off within the next three to five years.

The worldwide acceptance and use of standards such as Japan's i-mode and Europe's WAP, in combination with the work performed by market-based competition, collaboration of key-players, and regulations imposed by regulation authorities, are expected to boost consumer trust in m-commerce and strengthen its potential and perspectives.

REFERENCES

Adams, C. (2001). Mobile electronic payment systems: Main technologies and options. Retrieved August 9, 2004, from *http://www.bcs.org.uk/branches/hampshire/docs/mcommerce.ppt*

Cellular On-line. (2004). SIMPAY mobile payment platform announces first product. Retrieved August 11,2004, from *http://www.cellular.co.za/news_2004/feb/022704-simpay_mobile_payment_platform_a.htm*

Cellular On-line. (2004). Visa mobile 3D secure specification for m-commerce security. Retrieved August 10, 2004, from *http://www.cellular.co.za/technologies/mobile-3d/visa_mobile-3d.htm*

Centeno, C. (2002). Securing Internet payments: The potential of public key cryptography, public key infrastructure and digital signatures [*e*PSO background paper no.6]. Retrieved August 9, 2004, from *http://epso.jrc.es/backgrnd.html*

Charny, B. (2001). Nokia banks on mobile banking. *CNET News*. Retrieved August 9, 2004, from *http://news.com.com/2100-1033-276400.html?legacy=cnet&tag=mn_hd*

Clark, M. (2003). Government must regulate m-commerce. *Electric News Net*. Retrieved August 11, 2004, from *http://www.enn.ie/frontpage.news-9375556.html*

e-Europe Smart Card. (2003). Open smart card infrastructure for Europe, v2 , part 2-2: ePayments: Blueprint on mobile payments.*TB5 e/m Payment*. Retrieved August 12, 2004, from *http://www.eeurope-smartcards.org/Download/01-2-2.PDF*

EU Information Society Portal. (2004). e-Europe 2005, e-business. Retrieved August 12, 2004, from *http://europa.eu.int/information_society/eeurope/2005/all_about/mid_term_review/ebusiness/index_en.htm*

European Parliament (EP). (2000). Directive 2000/12/EC of the European Parliament and of the Council of 20 March 2000 relating to the taking up and pursuit of the business of credit institutions. *Official Journal, L 126*. Retrieved August 11, 2004, from *http://europa.eu.int/eur-lex/en*

European Parliament. (2000). Directive 2000/28/EC of the European Parliament and of the Council of 18 September 2000 amending Directive 2000/12/EC relating to the taking up and pursuit of the business of credit institutions. *Official Journal, L 275*. Retrieved August 11, 2004, from *http://europa.eu.int/eur-lex/en*

European Parliament. (2000). Directive 2000/46/EC of the European Parliament and of the Council of 18 September 2000 on the taking up, pursuit and prudential supervision of the business of electronic money institutions. *Official Journal, L 275*. Retrieved August 11,2004, from *http://europa.eu.int/eur-lex/en*

Nokia Press Releases. (2001). Nokia payment sSolution enables mobile e-commerce services with

multiple payment methods and enhanced security. Retrieved August 11,2004, from *http://press.nokia.com/PR/200102/809553_5.html*

NTT DoCoMo Web Site. (2004). I-mode. Retrieved August 10, 2004, from *http://www.nttdocomo.com/corebiz/imode*

Ryan, O. (2000). Japan's m-commerce boom. *BBC NEWS*. Retrieved August 10, 2004, from *http://news.bbc.co.uk/1/business/945051.stm*

T-Mobile Web Site. (2003). T-Mobile with CeBIT showcases on the subject of mobile commerce. Retrieved August 10, 2004, from *http://www.t-mobile-international.com/CDA/*

T-mobile_deutschland_newsdetails,1705,0,newsid-1787-yearid-1699-monthid-1755,en.html?w=736&h=435

Visa International Web Site. (2003). 3-D secure: System overview V.1.0.2 70015-01external version. Retrieved August 10, 2004, from *http://www.international.visa.com/fb/paytech/secure/pdfs/3DS_70015-01_System_Overview_external01_System_Overview_external_v1.0.2_May_2003.pdf*

Zika, J. (2004). Retail electronic money and prepaid payment instruments, thesis, Draft 1.4. Retrieved August 10, 2004, from *http://www.pay.czweb.org/en/Payment V1_4.pdf*

KEY TERMS

Bluetooth: A short-range radio technology aimed at simplifying communications among Internet devices and between devices and the Internet. It also aims to simplify data synchronization between Internet devices and other computers.

EMS: Enhanced Messaging Service. An application-level extension to SMS for cellular phones available on GSM, TDMA, and CDMA networks. An EMS-enabled mobile phone can send and receive messages that have special text formatting (i.e., bold or italic), animations, pictures, icons, sound effects, and special ringtones.

I-Mode: A wireless Internet service for mobile phones using HTTP, popular in Japan and increasingly elsewhere (i.e., USA, Germany, Belgium, France, Spain, Italy, Greece, Taiwan, etc.). It was inspired by WAP, which was developed in the U.S., and it was launched in 1999 in Japan. It became a runaway success because of its well-designed services and business model.

M-Commerce: Mobile commerce. Using mobile technology to access the Internet through a wireless device, such as a cell phone or a PDA, in order to sell or buy items (i.e., products or services), conduct a transaction, or perform supply chain or demand chain functions.

MMS: Multimedia Message Service. A store-and-forward method of transmitting graphics, video clips, sound files, and short text messages over wireless networks using the WAP protocol. It is based on multimedia messaging and is widely used in communication between mobile phones. It supports e-mail addressing without attachments.

MVNO: Mobile Virtual Network Operator. A company that does not own or control radio spectrum or associated radio infrastructure, but it does own and control its own subscriber base with the freedom to set tariffs and to provide enhanced value added services under its own brand.

PKI: Public Key Infrastructure. A system of digital certificates, certified authorities, and other registration authorities that verify and authenticate the validity of each party involved in an Internet transaction.

WAP: Wireless Application Protocol. A secure specification that allows users to access information instantly via handheld devices such as mobile phones, pagers, two-way radios, and so forth. It is supported by most wireless networks (i.e., GSM, CDMA, TETRA, etc.). WAP supports HTML and XML.

Exploiting Captions for Multimedia Data Mining

Neil C. Rowe
U.S. Naval Postgraduate School, USA

INTRODUCTION

Captions are text that describes some other information; they are especially useful for describing non-text media objects (images, audio, video, and software). Captions are valuable metadata for managing multimedia, since they help users better understand and remember (McAninch, Austin, & Derks, 1992-1993) and permit better indexing of media. Captions are essential for effective data mining of multimedia data, since only a small amount of text in typical documents with multimedia—1.2% in a survey of random World Wide Web pages (Rowe, 2002)—describes the media objects. Thus, standard Web browsers do poorly at finding media without knowledge of captions. Multimedia information is increasingly common in documents, as computer technology improves in speed and ability to handle it, and as people need multimedia for a variety of purposes like illustrating educational materials and preparing news stories.

Captions also are valuable, because non-text media rarely specify internally the creator, date, or spatial and temporal context, and cannot convey linguistic features like negation, tense, and indirect reference. Furthermore, experiments with users of multimedia retrieval systems show a wide range of needs (Sutcliffe et al., 1997) but a focus on media meaning rather than appearance (Armitage & Enser, 1997). This suggests that content analysis of media is unnecessary for many retrieval situations, which is fortunate, because it is often considerably slower and more unreliable than caption analysis. But using captions requires finding them and understanding them. Many captions are not clearly identified, and the mapping from captions to media objects is rarely easy. Nonetheless, the restricted semantics of media and captions can be exploited.

FINDING, RATING, AND INDEXING CAPTIONS

Background

Much text in a document near a media object is unrelated to that object, and text explicitly associated with an object often may not describe it (i.e., "JPEG picture here" or "Photo39573"). Thus, we need clues to distinguish and rate a variety of caption possibilities and words within them, allowing for more than one caption for an object or more than one object for a caption. Free commercial media search engines (i.e., images.google.com, multimedia.lycos.com, and www.altavista.com/image) use a few simple clues to index media, but their accuracy is significantly lower than that for indexing text. For instance, Rowe (2005) reported that none of five major image search engines could find pictures for "President greeting dignitaries" in 18 tries. So research is exploring a broader range of caption clues and types (Mukherjea & Cho, 1999; Sclaroff et al., 1999).

Sources of Captions

Some captions are explicitly attached to media objects by adding them to a digital library or database. On Web pages, HTML "alt" and "caption" tags also explicitly associate text with media objects. Clickable text links to media files are another good source of captions, since the text must explain the link. A short caption can be the name of the media file itself (e.g., "socket_wrench.gif").

Less explicit captions use conventions like centering or font changes to text. Titles and headings preceding a media object also can serve as captions, as they generalize over a block of information, but

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

they can be overly general. Paragraphs above, below, or next to media also can be captions, especially short paragraphs.

Other captions are embedded directly into the media, like characters drawn on an image (Lienhart & Wernicke, 2002) or explanatory words at the beginning of audio. These require specialized processing like optical character recognition to extract. Captions can be attached through a separate channel of video or audio, as with the "closed captions" associated with television broadcasts that aid hearing-impaired viewers and students learning languages. "Annotations" can function like captions, although they tend to emphasize analysis or background knowledge.

Cues for Rating Captions

A caption candidate's type affects its likelihood, but many other clues help rate it and its words (Rowe, 2005):

- Certain words are typical of captions, like those having to do with communication, representation, and showing. Words about space and time (e.g., "west." "event," "above," "yesterday") are good clues, too. Negative clues like "bytes" and "page" can be equally valuable as indicators of text unlikely to be captions. Words can be made to be more powerful clues by enforcing a limited or controlled vocabulary for describing media, like what librarians use in cataloging books (Arms, 1999), but this requires cooperation from caption writers and is often impossible.
- Position in the caption candidate matters: Words early in the text are four times more likely to describe a media object (Rowe, 2002).
- Distinctive phrases often signal captions (e.g., "the X above," "you can hear X," "X then Y") where X and Y describe depictable objects.
- Full parsing of caption candidates (Elworthy et al., 2001; Srihari & Zhang, 1999) can extract more detailed information about them, but it is time-consuming and prone to errors.
- Candidate length is a clue, since true captions average 200 characters with few under 20 or over 1,000.
- A good clue is words in common between the candidate caption and the name of the media file, such as "Front view of woodchuck burrowing" and image file "northern_woodchuck.gif."
- Nearness of the caption candidate to its media actually is not a clue (Rowe, 2002), since much nearby text in documents is unrelated.
- Some words in the name of a media file affect captionability (e.g., "view" and "clip" as positive clues and "icon" and "button" as negative clues).
- "Decorative" media objects occurring more than once on a page or three times on a site are 99% certain not to have captions (Rowe, 2002). Text generally captions only one media object except for headings and titles.
- Media-related clues are the size of the object (small objects are less likely to have captions) and the file format (e.g., JPEG images are more likely to have captions). Other clues are the number of colors and the ratio of width to length for an image.
- Consistency with the style of known captions on the same page or at the same site is also a clue because many organizations specify a consistent "look and feel" for their captions.

Quantifying Clues

Clue strength is the conditional probability of a caption given appearance of the clue, estimated from statistics by c/(c+n), where c is the number of occurrences of the clue in a caption and n is the number of occurrences of the clue in a noncaption. If we have a representative sample, clue appearances can be modeled as a binomial process with expected standard deviation $\sqrt{cn/(c+n)}$. This can be used to judge whether a clue is statistically significant, and it rules out many potential word clues. Recall-precision analysis also can compare clues; Rowe (2002) showed that text-word clues were the most valuable in identifying captions, followed in order by caption type, image format, words in common between the text and the image filename, image size, use of digits in the image file name, and image-filename word clues.

Methods of data mining (Witten & Frank, 2000) can combine clues to get an overall likelihood that

some text is a caption. Linear models, Naive-Bayes models, and case-based reasoning have been used. The words of the captions can be indexed, and the likelihoods can be used by a browser to sort media for presentation to the user that match a set of keywords.

MAPPING CAPTIONS TO MULTIMEDIA

Background

Studies show that users usually consider media data as "depicting" a set of objects (Jorgensen, 1998) rather than a set of textures arranged in space or time. Captions can be:

- **Component-Depictive:** The caption describes objects and/or processes that correspond to particular parts of the media. For instance, a caption "President speaking to board" with a picture that shows a president behind a podium with several other people. This caption type is quite common.
- **Whole-Depictive:** The caption describes the media as a whole. This is often signaled by media-type words like "view," "clip." and "recording"; for instance, "Tape of City Council 7/26/04" with some audio. Such captions summarize overall characteristics of the media object and help distinguish it from others. Adjectives are especially helpful, as in "infrared picture." "short clip," and "noisy recording"; they specify distributions of values. Dates and locations for associated media can be found in special linguistic formulas (Smith, 2002).
- **Illustrative-Example:** The media presents only an example of the phenomenon described by the caption; for instance, "War in the Gulf" with a picture of tanks in a desert.
- **Metaphorical:** The media represents something related to the caption but does not depict it or describe it; for instance, "Military fiction" with a picture of tanks in a desert.
- **Background:** The caption only gives background information about the media; for instance, "World War II" with a picture of Winston Churchill. *National Geographic* magazine often uses caption sentences of this kind after the first sentence.

Media Properties and Structure

The structure of media objects can be referenced by component-depictive caption sentences to orient the viewer or listener. Then valuable information is often contained in the sub-objects of a media object that captions do not convey. Images, audio, and video are multidimensional signals for which local changes in the signal characteristics help segment them into sub-objects (Aslandogan & Yu, 1999). Color or texture changes in an image suggest separate objects; changes in the frequency-intensity plot of audio suggest beginnings and ends of sounds; and many simultaneous changes between corresponding locations in two video frames suggest a new shot (Wactlar et al, 2000). But segmentation methods are not especially reliable. Also, some media objects have multiple colors or textures, like images of trees or human faces, and domain-dependent knowledge must group regions into larger objects.

Software can calculate properties of segmented regions and classify them. Mezaris, Compatsiaris, and Strinzis (2003), for instance, classify image regions by color, size, shape, and relative position, and then infer probabilities for what they could represent. Additional laws of media space can rule out possibilities so that objects closer to a camera appear larger, and gravity is downward, so support relationships between objects often can be found (e.g., people on floors). Similarly, the pattern of voices and the duration of their speaking times in an audio recording can suggest in general terms what is happening. The subject of a media object often can be inferred, even without a caption, since subjects are typically near the center of the media space, not touching its edges, and well distinguished from nearby regions in intensity or texture.

Caption-Media Correspondence

While finding the caption-media correspondence for component-depictive captions can be generally difficult, there are easier subcases. One is the recognition and naming of faces in an image (Satoh, Nakamura, & Kanda, 1999). Another is captioned graphics, since their structure is easier to infer than most images (Preim et al., 1998).

In general, grammatical subjects of a caption often correspond to the principal subjects within the

media (Rowe, 2005). For instance, "Large deer beside tree" has the grammatical subject "deer," and we would expect to see all of it in the picture near the center, whereas "tree" has no such guarantee. Exceptions are undepictable abstract subjects (i.e., "Jobless rate soars"). Present-tense principal verbs and verbals can depict dynamic physical processes, such as "eating" in "Deer eating flowers," and direct objects of such verbs and verbals usually are fully depicted in the media when they are physical like "flowers." Objects of physical-location prepositions attached to the principal subject are also depicted in part (but not necessarily as a whole). Subjects that are media objects like "view" defer viewability to their objects. Motion-denoting words can be depicted directly in video, audio, and software, rather than just their subjects and objects. They can be translational (e.g. "go"), configurational ("develop"), property-changing ("lighten"), relationship-changing ("fall"), social ("report"), or existential ("appear").

Captions are "deictic," using the linguistic term for expressions whose meaning requires assimilation of information from outside the expression itself. Spatial deixis refers to spatial relationships between objects or parts of objects and entails a set of physical constraints (DiTomaso et al., 1998; Pineda & Garza, 2000). Spatial deixis expressions like "above" and "outside" are often "fuzzy" in that they do not define a precise area but rather associate a probability distribution with a region of space (Matsakis et al., 2001). It is important to determine the reference location of the referring expression, which is usually the characters of the text itself but can be previously referenced objects like "right" in "the right picture below." Some elegant theory has been developed, although captions on media objects that use such expressions are not especially common.

Media objects also can occur in sets with intrinsic meaning. The media can be a time sequence, a causal sequence, a dispersion in physical space, or a hierarchy of concepts. Special issues arise when captions serve to distinguish one media object from another (Heidorn, 1999). Media-object sets also can be embedded in other sets. Rules for set correspondences can be learned from examples (Cohen, Wang, & Murphy, 2003).

For deeper understanding of media, the words of the caption can be matched to regions of the media. This permits applications like calculating the size and contrast of media subobjects mentioned in the caption, recognizing the time of day when it is not mentioned, and recognizing additional unmentioned objects. Matching must take into account the properties of the words and regions, and the constraints relating them, and must try to find the best matches. Statistical methods similar to those for identifying clues for captions can be used, except that there are many more categories, entailing problems of obtaining enough data. Some help is provided by knowledge of the settings of things described in captions (Sproat, 2001). Machine learning methods can learn the associations between words and types of image regions (Barnard et al., 2003; Roy, 2000, 2001).

Generating Captions

Since captions are so valuable in indexing and explaining media objects, it is important to obtain good ones. The methods described above for finding caption candidates can be used to collect text for a caption when an explicit one is lacking. Media content analysis also can provide information that can be paraphrased into a caption; this is most possible with graphics images. Discourse theory can help to make captions sound natural by providing "discourse strategies" such as organizing the caption around one media attribute that determines all the others (e.g., the department in a budget diagram) (Mittal et al., 1998). Then guidelines about how much detail the user wants, together with a ranking of the importance of specific details, can be used to assemble a reasonable set of details to mention in a caption. Semi-automated techniques also can construct captions by allowing users to point and click within media objects and supply audio (Srihari & Zhang, 2000). Captions also can be made "interactive" so that changes to them cause changes in corresponding media (Preim et al., 1998).

FUTURE TRENDS

Future multimedia-retrieval technology will not be dramatically different, although multimedia will be increasingly common in many applications. Captions will continue to provide the easiest access via keyword search, and caption text will remain important to explain media objects in documents. But improved

media content analysis (aided by speed increases in computer hardware) will increasingly help in both disambiguating captions and mapping their words to parts of the media object. Machine-learning methods will be used increasingly to learn the necessary associations.

CONCLUSION

Captions are essential tools to managing and manipulating multimedia objects as one of the most powerful forms of metadata. A good multimedia data-mining system needs to include captions and their management in its design. This includes methods for finding them in unrestricted text as well as ways of mapping them to the media objects. With good support for captions, media objects are much better integrated with the traditional text data used by information systems.

REFERENCES

Armitage, L.H., & Enser, P. (1997). Analysis of user need in image archives. *Journal of Information Science, 23*(4), 287-299.

Arms, L. (1999). Getting the picture: Observations from the Library of Congress on providing access to pictorial images. *Library Trends, 48*(2), 379-409.

Aslandogan, Y., & Yu, C. (1999). Techniques and systems for image and video retrieval. *IEEE Transactions on Knowledge and Data Engineering, 11*(1), 56-63.

Barnard, K., et al. (2003). Matching words and pictures. *Journal of Machine Learning Research, 3*, 1107-1135.

Cohen, W., Wang, R., & Murphy, R. (2003). Understanding captions in biomedical publications. *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, Washington, D.C.

DiTomaso, V., Lombardo, V., & Lesmo, L. (1998). A computational model for the interpretation of static locative expressions. In P. Oliver, & K.-P. Gapp (Eds.), *Representation and processing of spatial expressions* (pp. 73-90). Mahwah, NJ: Lawrence Erlbaum.

Elworthy, D., Rose, T., Clare, A., & Kotcheff, A. (2001). A natural language system for retrieval of captioned images. *Natural Language Engineering, 7*(2), 117-142.

Heidorn, P.B. (1999). The identification of index terms in natural language objects. *Proceedings of the Annual Conference of the American Society for Information Science*, Washington, D.C.

Jorgensen, C. (1998). Attributes of images in describing tasks. *Information Processing and Management, 34*(2/3), 161-174.

Lienhart, R., & Wernicke, A. (2002). Localizing and segmenting text in video, images, and Web pages. *IEEE Transactions on Circuits and Systems for Video Technology, 12*(4), 256-268.

Matsakis, P., Keller, J., Wendling, L., Marjarnaa, & Sjahputera, O. (2001). Linguistic description of relative positions in images. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, 31*(4), 573-588.

McAninch, C., Austin, J., & Derks, P. (1992-1993). Effect of caption meaning on memory for nonsense figures. *Current Psychology Research & Reviews, 11*(4), 315-323.

Mezaris, V., Kompatsiaris, I., & Strinzis, M. (2003). An ontology approach to object-based image retrieval. *Proceedings of the International Conference on Image Processing*, Barcelona, Spain, (Vol. 2, pp. 511-514).

Mittal, V., Moore, J., Carenini, J., & Roth, S. (1998). Describing complex charts in natural language: A caption generation system. *Computational Linguistics, 24*(3), 437-467.

Mukherjea, S., & Cho, J. (1999). Automatically determining semantics for World Wide Web multimedia information retrieval. *Journal of Visual Languages and Computing, 10*, 585-606.

Pineda, L., & Garza, G. (2000). A model for multimodal reference resolution. *Computational Linguistics, 26*(2), 139-193.

Preim, B., Michel, R., Hartmann, K., & Strothotte, T. (1998). Figure captions in visual interfaces. *Proceedings of the Working Conference on Advanced Visual Interfaces*, L'Aquila, Italy.

Rowe, N. (2002). MARIE-4: A high-recall, self-improving Web crawler that finds images using captions. *IEEE Intelligent Systems*, *17*(4), 8-14.

Rowe, N. (2005). Exploiting captions for Web data mining. In A. Scime (Ed.), *Web mining: Applications and techniques* (pp. 119-144). Hershey, PA: Idea Group Publishing.

Roy, D.K. (2000/2001). Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication, 4*(1), 33-56.

Satoh, S., Nakamura, Y., & Kanda, T. (1999). Name-it: Naming and detecting faces in news videos. *IEEE Multimedia, 6*(1), 22-35.

Sclaroff, S., La Cascia, M., Sethi, S., & Taycher, L. (1999). Unifying textual and visual cues for content-based image retrieval on the World Wide Web. *Computer Vision and Image Understanding*, *75*(1/2), 86-98.

Smith, D. (2002). Detecting events with date and place information in unstructured texts. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, Portland, Oregon.

Sproat, R. (2001). Inferring the environment in a text-to-scene conversion system. *Proceedings of the International Conference on Knowledge Capture*, Victoria, British Columbia, Canada.

Srihari, R., & Zhang, Z. (1999). Exploiting multimodal context in image retrieval. *Library Trends, 48(*2), 496-520.

Srihari, R., & Zhang, Z. (2000). Show&Tell: A semi-automated image annotation system. *IEEE Multimedia, 7*(3), 61-71.

Sutcliffe, A., Hare, M., Doubleday, A., & Ryan, M. (1997). Empirical studies in multimedia information retrieval. In M. Maybury (Ed.), *Intelligent multimedia information retrieval* (pp. 449-472). Menlo Park, CA: AAAI Press/MIT Press.

Wactlar, H., Hauptmann, A., Christel, M., Houghton, R., & Olligschlaeger, A. (2000). Complementary video and audio analysis for broadcast news archives. *Communications of the ACM, 43*(2), 42-47.

Witten, I., & Frank, E. (2000). *Data mining: Practical machine learning with Java implementations*. San Francisco, CA: Morgan Kaufmann.

KEY TERMS

"Alt" String: An HTML tag for attaching text to a media object.

Caption: Text describing a media object.

Controlled Vocabulary: A limited menu of words from which metadata like captions must be constructed.

Data Mining: Searching for insights in large quantities of data.

Deixis: A linguistic expression whose understanding requires understanding something besides itself, as with a caption.

HTML: Hypertext Markup Language, the base language of pages on the World Wide Web.

Media Search Engine: A Web search engine designed to find media (usually images) on the Web.

Metadata: Information describing another data object such as its size, format, or description.

Web Search Engine: A Web site that finds other Web sites whose contents match a set of keywords, using a large index to Web pages.

Face for Interface

Maja Pantic
Delft University of Technology, The Netherlands

INTRODUCTION: THE HUMAN FACE

The human face is involved in an impressive variety of different activities. It houses the majority of our sensory apparatus—eyes, ears, mouth, and nose—allowing the bearer to see, hear, taste, and smell. Apart from these biological functions, the human face provides a number of signals essential for interpersonal communication in our social life. The face houses the speech production apparatus and is used to identify other members of the species; it regulates conversation by gazing or nodding and interprets what has been said by lip reading. It is our direct and naturally preeminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression (Lewis & Haviland-Jones, 2000). Personality, attractiveness, age, and gender also can be seen from someone's face. Thus, the face is a multisignal sender/receiver capable of tremendous flexibility and specificity. In general, the face conveys information via four kinds of signals listed in Table 1.

Automating the analysis of facial signals, especially rapid facial signals, would be highly beneficial for fields as diverse as security, behavioral science, medicine, communication, and education. In security contexts, facial expressions play a crucial role in establishing or detracting from credibility. In medicine, facial expressions are the direct means to identify when specific mental processes are occurring. In education, pupils' facial expressions inform the teacher of the need to adjust the instructional message.

As far as natural interfaces between humans and computers (i.e., PCs, robots, machines) are concerned, facial expressions provide a way to communicate basic information about needs and demands to the machine. In fact, automatic analysis of rapid facial signals seems to have a natural place in various vision subsystems, including automated tools for gaze and focus of attention tracking, lip reading, bimodal speech processing, face/visual speech synthesis, face-based command issuing, and facial affect processing. Where the user is looking (i.e., gaze tracking) can be effectively used to free computer users from the classic keyboard and mouse. Also, certain facial signals (e.g., a wink) can be associated with certain commands (e.g., a mouse click), offering an alternative to traditional keyboard and mouse commands. The human capability to hear in noisy environments by means of lip reading is the basis for bimodal (audiovisual) speech processing that can lead to the realization of robust speech-driven interfaces. To make a believable talking head (avatar) representing a real person, tracking the person's facial signals and making the avatar mimic those

Table 1. Four types of facial signals

• *Static facial signals* represent relatively permanent features of the face, such as the bony structure, the soft tissue, and the overall proportions of the face. These signals are usually exploited for person identification.
• *Slow facial signals* represent changes in the appearance of the face that occur gradually over time, such as development of permanent wrinkles and changes in skin texture. These signals can be used for assessing the age of an individual.
• *Artificial signals* are exogenous features of the face such as glasses and cosmetics. These signals provide additional information that can be used for gender recognition.
• *Rapid facial signals* represent temporal changes in neuromuscular activity that may lead to visually detectable changes in facial appearance, including blushing and tears. These (atomic facial) signals underlie *facial expressions*.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 2. Examples of facial action units (AUs)

	AU1: Raised inner eyebrow		AU2: Raised outer eyebrow
	AU1 + AU2: Raised eyebrows		AU4: Lowered eyebrow Eyebrows drawn together
	AU5: Raised upper eyelid		AU6: Raised cheek Compressed eyelid
	AU7: Tightened eyelid		AU41: Drooped eyelid
	AU44: Squinted eyes		AU46: Wink
	AU9: Wrinkled nose		AU11: Deepened nasolabial furrow
	AU12: Lip corners pulled up		AU13: Lip corners pulled up sharply
	AU14: Dimpler – mouth corners pulled inwards		AU15: Lip corners depressed
	AU17: Chin raised		AU19: Tongue shown
	AU20: Mouth stretched horizontally		AU24: Lips pressed
	AU26: Jaw dropped		AU29: Jaw pushed forward
	AU30: Jaw sideways		AU36: Bulge produced by the tongue

using synthesized speech and facial expressions are compulsory. The human ability to read emotions from someone's facial expressions is the basis of facial affect processing that can lead to expanding interfaces with emotional communication and, in turn, obtain a more flexible, adaptable, and natural interaction between humans and machines.

It is this wide range of principle driving applications that has lent a special impetus to the research problem of automatic facial expression analysis and produced a surge of interest in this research topic.

BACKGROUND: FACIAL ACTION CODING

Rapid facial signals are movements of the facial muscles that pull the skin, causing a temporary distortion of the shape of the facial features and of the appearance of folds, furrows, and bulges of skin. The common terminology for describing rapid facial signals refers either to culturally dependent linguistic terms, indicating a specific change in the appearance of a particular facial feature (e.g., smile, smirk, frown, sneer), or for linguistic universals describing the activity of specific facial muscles that caused the observed facial appearance changes.

There are several methods for linguistically universal recognition of facial changes based on the facial muscular activity (Scherer & Ekman, 1982). From those, the facial action coding system (FACS) proposed by Ekman et al. (1978, 2002) is the best-known and most commonly used system. It is a system designed for human observers to describe changes in the facial expression in terms of visually observable activations of facial muscles. The changes in the facial expression are described with FACS in terms of 44 different Action Units (AUs), each of which is anatomically related to the contraction of either a specific facial muscle or a set of facial muscles. Examples of different AUs are given in Table 2. Along with the definition of various AUs,

FACS also provides the rules for visual detection of AUs and their temporal segments (i.e., onset, apex, offset) in a face image. Using these rules, a FACS coder (i.e., a human expert having formal training in using FACS) decomposes a shown facial expression into the AUs that produce the expression.

Although FACS provides a good foundation for AU coding of face images by human observers, achieving AU recognition by a computer is by no means a trivial task. A problematic issue is that AUs can occur in more than 7,000 different complex combinations (Scherer & Ekman, 1982), causing bulges (e.g., by the tongue pushed under one of the lips) and various in- and out-of-image-plane movements of permanent facial features (e.g., jetted jaw) that are difficult to detect in 2D face images.

AUTOMATED FACIAL ACTION CODING

Most approaches to automatic facial expression analysis attempt to recognize a small set of prototypic emotional facial expressions (i.e., fear, sadness, disgust, anger, surprise, and happiness) (for an exhaustive survey of the past work on this research topic, the reader is referred to the work of Pantic & Rothkrantz [2003]). This practice may follow from the work of Darwin and more recently Ekman (Lewis & Haviland-Jones, 2000), who suggested that basic emotions have corresponding prototypic expressions. In everyday life, however, such prototypic expressions occur relatively rarely; emotions are displayed more often by subtle changes in one or few discrete facial features such as raising the eyebrows in surprise. To detect such subtlety of human emotions and, in general, to make the information conveyed by facial expressions available for usage in the various applications mentioned above, automatic recognition of rapid facial signals (AUs) is needed.

Few approaches have been reported for automatic recognition of AUs in images of faces. Some researchers described patterns of facial motion that correspond to a few specific AUs, but did not report on actual recognition of these AUs. Examples of such works are the studies of Mase (1991) and Essa and Pentland (1997). Almost all other efforts in automating FACS coding addressed the problem of automatic AU recognition in face video using both machine vision techniques like optical flow analysis, Gabor wavelets, temporal templates, particle filtering, and machine learning techniques such as neural networks, support vector machines, and hidden Markov models. To detect six individual AUs in face image sequences free of head motions, Bartlett et al. (1999) used a neural network. They achieved 91% accuracy by feeding the pertinent network with the results of a hybrid system combining holistic spatial analysis and optical flow with local feature analysis. To recognize eight individual AUs and four combinations of AUs with an average recognition rate of 95.5% for face image sequences free of head motions, Donato et al. (1999) used Gabor wavelet representation and independent component analysis. To recognize eight individual AUs and seven combinations of AUs with an average recognition rate of 85% for face image sequences free of head motions, Cohn et al. (1999) used facial feature point tracking and discriminant function analysis. Tian et al. (2001) used lip tracking, template matching, and neural networks to recognize 16 AUs occurring alone or in combination in nearly frontal-view face image sequences. They reported an 87.9% average recognition rate attained by their method. Braathen et al. (2002) reported on automatic recognition of three AUs using particle filtering for 3D tracking, Gabor wavelets, support vector machines, and hidden Markov models to analyze an input face image sequence having no restriction placed on the head pose. To recognize 15 AUs occurring alone or in combination in a nearly frontal-view face image sequence, Valstar et al. (2004) used temporal templates. Temporal templates are 2D images constructed from image sequences, which show where and when motion in the image sequence has occurred. The authors reported a 76.2% average recognition rate attained by their method.

In contrast to all these approaches to automatic AU detection, which deal only with frontal-view face images and cannot handle temporal dynamics of AUs, Pantic and Patras (2004) addressed the problem of automatic detection of AUs and their temporal segments (onset, apex, offset) from profile-view face image sequences. They used particle filtering to track 15 fiducial facial points in an input face-profile video and temporal rules to recognize temporal segments of 23 AUs occurring alone or in

a combination in the input video sequence. They achieved an 88% average recognition rate by their method.

The only work reported to date that addresses automatic AU coding from static face images is the work of Pantic and Rothkrantz (2004). It concerns an automated system for AU recognition in static frontal- and/or profile-view color face images. The system utilizes a multi-detector approach for facial component localization and a rule-based approach for recognition of 32 individual AUs. A recognition rate of 86% is achieved by the method.

CRITICAL ISSUES

Facial expression is an important variable for a large number of basic science studies (in behavioral science, psychology, psychophysiology, psychiatry) and computer science studies (in natural human-machine interaction, ambient intelligence, affective computing). While motion records are necessary for studying temporal dynamics of facial behavior, static images are important for obtaining configurational information about facial expressions, which is essential, in turn, for inferring the related meaning (i.e., in terms of emotions) (Scherer & Ekman, 1982). As can be seen from the survey given above, while several efforts in automating FACS coding from face video have been made, only Pantic and Rothkrantz (2004) made an effort for the case of static face images.

In a frontal-view face image (portrait), facial gestures such as showing the tongue (AU 19) or pushing the jaw forwards (AU 29) represent out-of-image-plane, non-rigid facial movements that are difficult to detect. Such facial gestures are clearly observable in a profile view of the face. Hence, the usage of face-profile view promises a qualitative enhancement of AU detection performed by enabling detection of AUs that are difficult to encode in a frontal facial view. Furthermore, automatic analysis of expressions from face profile-view would facilitate deeper research on human emotion. Namely, it seems that negative emotions (where facial displays of AU2, AU4, AU9, and the like are often involved) are more easily perceivable from the left hemiface than from the right hemiface, and that, in general, the left hemiface is perceived to display more emotion than the right hemiface (Mendolia & Kleck, 1991). However, only Pantic and Patras (2004) made an effort to date to automate FACS coding from video of profile faces. Finally, it seems that facial actions involved in spontaneous emotional expressions are more symmetrical, involving both the left and the right side of the face, than deliberate actions displayed on request. Based upon these observations, Mitra and Liu (2004) have shown that facial asymmetry has sufficient discriminating power to significantly improve the performance of an automated genuine emotion classifier. In summary, the usage of both frontal and profile facial views and moving toward 3D analysis of facial expressions promises, therefore, a qualitative increase in facial behavior analysis that can be achieved. Nevertheless, only Braathen et al. (2002) made an effort to date in automating FACS coding using a 3D face representation.

There is now a growing body of psychological research that argues that temporal dynamics of facial behavior (i.e., timing, duration, and intensity of facial activity) is a critical factor for the interpretation of observed behavior (Lewis & Haviland-Jones, 2000). For example, Schmidt and Cohn (2001) have shown that spontaneous smiles, in contrast to posed smiles, are fast in onset, can have multiple AU12 apexes (i.e., multiple rises of the mouth corners), and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within one second. Hence, it is obvious that automated tools for the detection of AUs and their temporal dynamics would be highly beneficial. However, only Pantic and Patras (2004) reported so far on an effort to automate the detection of the temporal segments of AUs in face image sequences.

None of the existing systems for facial action coding in images of faces is capable of detecting all 44 AUs defined by the FACS system. Besides, in many instances strong assumptions are made to make the problem more tractable (e.g., images contain faces with no facial hair or glasses, the illumination is constant, the subjects are young and of the same ethnicity). Only the method of Braathen et al. (2002) deals with rigid head motions, and only the method of Essa and Pentland (1997) can handle distractions like facial hair (i.e., beard, moustache) and glasses. None of the automated facial expression analyzers proposed in the literature to date fills

in missing parts of the observed face; that is, none perceives a whole face when a part of it is occluded (i.e., by a hand or some other object). Also, though the conclusions generated by an automated facial expression analyzer are affected by input data certainty, robustness of the applied processing mechanisms, and so forth, except for the system proposed by Pantic and Rothkrantz (2004), no existing system for automatic facial expression analysis calculates the output data certainty.

In spite of repeated references to the need for a readily accessible reference set of static images and image sequences of faces that could provide a basis for benchmarks for efforts in automating FACS coding, no database of images exists that is shared by all diverse facial-expression-research communities. In general, only isolated pieces of such a facial database exist. An example is the unpublished database of Ekman-Hager Facial Action Exemplars. It has been used by Bartlett et al. (1999), Donato et al. (1999), and Tian et al. (2001) to train and test their methods for AU detection from face image sequences. The facial database made publicly available, but still not used by all diverse facial-expression-research communities, is the Cohn-Kanade AU-coded Face Expression Image Database (Kanade et al., 2000). None of these databases contains images of faces in profile view, none contains images of all possible single-AU activations, and none contains images of spontaneous facial expressions. Also, the metadata associated with each database object usually does not identify the temporal segments of AUs shown in the face video in question. This lack of suitable and common training and testing material forms the major impediment to comparing, resolving, and extending the issues concerned with facial micro-action detection from face video. It is, therefore, a critical issue that should be addressed in the nearest possible future.

CONCLUSION

Faces are tangible projector panels of the mechanisms that govern our emotional and social behaviors. Analysis of facial expressions in terms of rapid facial signals (i.e., in terms of the activity of the facial muscles causing the visible changes in facial expression) is, therefore, a highly intriguing problem. While the automation of the entire process of facial action coding from digitized images would be enormously beneficial for fields as diverse as medicine, law, communication, education, and computing, we should recognize the likelihood that such a goal still belongs to the future. The critical issues concern the establishment of basic understanding of how to achieve automatic spatio-temporal facial-gesture analysis from multiple views of the human face and the establishment of a readily accessible centralized repository of face images that could provide a basis for benchmarks for efforts in the field.

REFERENCES

Bartlett, M.S., Hager, J.C., Ekman, P., & Sejnowski, T.J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology, 36*, 253-263.

Braathen, B., Bartlett, M.S., Littlewort, G., Smith, E., & Movellan, J.R. (2002). An approach to automatic recognition of spontaneous facial actions. *Proceedings of the International Conference on Face and Gesture Recognition (FGR'02)*, Washington, USA, (pp. 345-350).

Cohn, J.F., Zlochower, A.J., Lien, J., & Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual faces coding. *Psychophysiology, 36*, 35-43.

Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., & Sejnowski, T.J. (1999). Classifying facial actions. *IEEE Trans. Pattern Analysis and Machine Intelligence, 21*(10), 974-989.

Ekman, P., & Friesen, W.V. (1978). *Facial action coding system*. Palo Alto, CA: Consulting Psychologist Press.

Ekman, P., Friesen, W.V., & Hager, J.C. (2002). *Facial action coding system.* Salt Lake City, UT: Human Face.

Essa, I., & Pentland, A. (1997). Coding, analysis, interpretation and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence, 19*(7), 757-763.

Kanade, T., Cohn, J., & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proceedings of the International Conference on Face and Gesture Recognition*, Grenoble, France, (pp. 46-53).

Lewis, M., & Haviland-Jones, J.M. (Eds.). (2000). *Handbook of emotions*. New York: Guilford Press.

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Transactions, E74*(10), 3474-3483.

Mendolia, M., & Kleck, R.E. (1991). Watching people talk about their emotions—Inferences in response to full-face vs. profile expressions. *Motivation and Emotion, 15*(4), 229-242.

Mitra, S., & Liu, Y. (2004). Local facial asymmetry for expression classification. *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Washington, USA, (pp. 889-894).

Pantic, M., & Patras, I. (2004). Temporal modeling of facial actions from face profile image sequences. *Proceedings of the International Conference on Multimedia and Expo.*, Taipei, Taiwan, (Vol. 1, pp. 49-52).

Pantic, M., & Rothkrantz, L.J.M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *IEEE, 91*(9), 1370-1390.

Pantic, M., & Rothkrantz, L.J.M. (2004). Facial action recognition for facial expression analysis from static face images. *IEEE Trans. Systems, Man, and Cybernetics – Part B, 34*(3), 1449-1461.

Scherer, K.R., & Ekman, P. (Eds.). (1982). *Handbook of methods in non-verbal behavior research*. Cambridge, MA: Cambridge University Press.

Schmidt, K.L., & Cohn, J.F. (2001). Dynamics of facial expression: Normative characteristics and individual differences. *Proceedings of the International Conference on Multimedia and Expo.*, Tokyo, Japan, (pp. 547-550).

Tian, Y., Kanade, T., & Cohn, J.F. (2001). Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence, 23*(2), 97-115.

Valstar, M.F., Patras, I., & Pantic, M. (2004). Facial action unit recognition using temporal templates. *Proceedings of the International Workshop on Robot-Human Interaction,* Kurashiki, Japan, (pp. 253-258).

F

KEY TERMS

Ambient Intelligence: The merging of mobile communications and sensing technologies with the aim of enabling a pervasive and unobtrusive intelligence in the surrounding environment supporting the activities and interactions of the users. Technologies like face-based interfaces and affective computing are inherent ambient-intelligence technologies.

Automatic Facial Expression Analysis: A process of locating the face in an input image, extracting facial features from the detected face region, and classifying these data into some facial-expression-interpretative categories such as facial muscle action categories, emotion (affect), attitude, and so forth.

Face-Based Interface: Regulating (at least partially) the command flow that streams between the user and the computer by means of facial signals. This means associating certain commands (e.g., mouse pointing, mose clicking, etc.) with certain facial signals (e.g., gaze direction, winking, etc.). Face-based interface can be effectively used to free computer users from classic keyboard and mouse commands.

Face Synthesis: A process of creating a talking head that is able to speak, display (appropriate) lip movements during speech, and display expressive facial movements.

Lip Reading: The human ability to hear in noisy environments by analyzing visible speech signals; that is, by analyzing the movements of the lips and the surrounding facial region. Integrating both visual speech processing and acoustic speech processing results in a more robust bimodal (audiovisual) speech processing.

Machine Learning: A field of computer science concerned with the question of how to construct computer programs that automatically im-

prove with experience. The key algorithms that form the core of machine learning include neural networks, genetic algorithms, support vector machines, Bayesian networks, and Markov models.

Machine Vision: A field of computer science concerned with the question of how to construct computer programs that automatically analyze images and produce descriptions of what is imaged.

FDD Techniques Towards the Multimedia Era

F

Athanassios C. Iossifides
COSMOTE S.A., Greece

Spiros Louvros
COSMOTE S.A., Greece

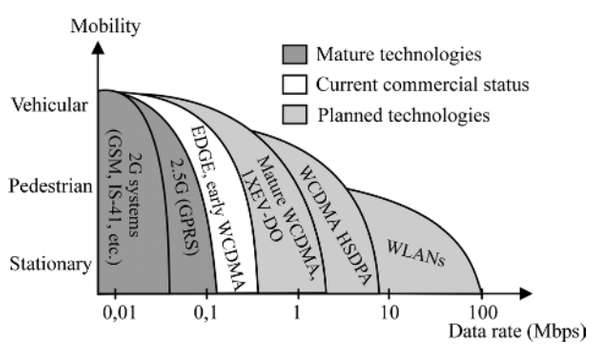
Stavros A. Kotsopoulos
University of Patras, Greece

INTRODUCTION

Global rendering of personalized multimedia services is the key issue determining the evolution of next-generation mobile networks. The determinant factor of mobile multimedia communications feasibility is the air-interface technology. The Universal Mobile Telecommunications System (UMTS) evolution, based on wideband code-division multiple access (WCDMA), constitutes a major step to the target of truly ubiquitous computing: computing anywhere, anytime, guaranteeing mobility and transparency. However, certain steps are still required in order to achieve the desired data rates, capacity, and quality of service (QoS) of different traffic classes inherent in multimedia services.

A view of data-rate trends of applied and future mobile communications technologies is shown in Figure 1. UMTS, being in its premature application stage, is currently providing rates up to 64/384 Kbps (uplink [UL]/downlink [DL]). It was initially designed to provide rates up to 2 Mbps under ideal conditions, which seems not enough from a competitiveness point of view compared to WLANs (wireless local-area networks) that aim to easily reach 2- to 10-Mbps data rates with the possibility of reaching 100 Mbps (Simoens, Pellati, Gosteau, Gosse, & Ware, 2003). Hardware, software, installation, and operational costs of 3G (3rd Generation) systems could be proven unjustified and unprofitable if they cannot cope with at least a certain share of data rates over 2 Mbps. This article focuses on the characteristics, application, and future enhancements (planned in 3GPP Release 5 and 6 or under research) of WCDMA-FDD (frequency-division duplex) toward high-quality multimedia services.

Figure 1. Data-rate trends of mobile communications technologies (see also Honkasalo, Pehkonen, Niemi, & Leino, 2002)



GSM: Global system for mobile communications
GPRS: General packet radio services
EDGE: Enhanced data for GSM evolution
1XEV-DO: CDMA 1X evolution - data only
HSDPA: High speed downlink packet access

CDMA BACKGROUND

CDMA, in contrast to FDMA (Frequency Division Multiple Access) and TDMA (Time Division Multiple Access), poses no restrictions to the time interval and frequency band to be used for the transmission of different users. All users can transmit simultaneously while occupying the whole available bandwidth (Figure 2). They are separated by uniquely (per user) assigned codes with proper low cross-interference properties. Thus, while interuser interference is strictly avoided in TDMA and FDMA systems by assigning different portions of time (time slots [TSs]) or bandwidth to different users, respectively, interuser interference, referred to as multiple-access interference (MAI), is inherent in CDMA techniques and is the limiting capacity factor (interference-limited systems).

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 2. FDMA, TDMA, and CDMA principles

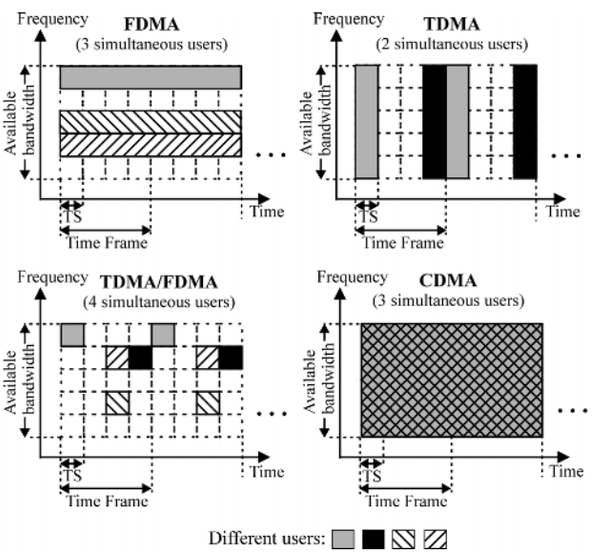
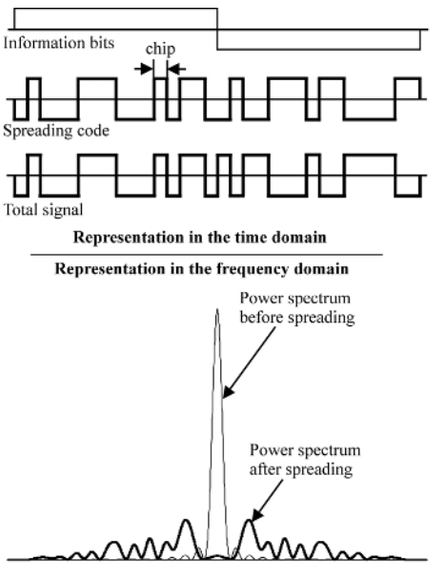


Figure 3. DS/CDMA principle



Although CDMA has been known for several decades, only in the last two decades has interest peaked regarding its use for mobile communications because of its enhanced performance compared to standard TDMA and FDMA techniques. Greater capacity, exploitation of multipath fading through RAKE combining, soft handover, and soft capacity are some of CDMA's advantages (Viterbi, 1995). The first commercial CDMA mobile application was IS-95 (1993). The real boost of CDMA applications, though, was the adoption of the WCDMA air interface for UMTS.

CDMA is applied using spread-spectrum techniques, such as frequency hopping (FH), direct sequence (DS), or hybrid methods. The DS technique, which is used in UMTS, is applied by multiplying the information symbols with faster pseudorandom codes of low cross-correlation between each other, which spreads the information bandwidth (Figure 3). The number of code pulses (chips) used for spreading an information symbol is called the spreading factor (SF). The higher the SF, the greater the tolerance to MAI is. A simplified block diagram of a CDMA transmitter and receiver is given in Figure 4. The receiver despreads the signal with the specific user's unique code followed by an integrator or digital summing device. Coexistent users' signals act as additive wideband noise (MAI).

With properly selected codes (of low autocorrelation), multipath propagation can turn into diversity gain for CDMA systems as soon as multiple paths' delays are spaced more than the chip duration (these paths are called resolved). In such a case, a RAKE receiver is employed (Figure 5), which performs a full reception procedure for each one of the resolved paths and properly combines the received signal replicas. In any case, discrimination between CDMA users is feasible with conventional receivers (no multiuser receivers) only when an advanced power-control method is engaged. Otherwise the near-far effect will destroy multiple-access capability.

There is no universally accepted definition for what is called WCDMA. From a theoretical point of view, a CDMA system is defined as wideband when the overall spread bandwidth exceeds the coherence bandwidth of the channel (Milstein, 2000). In such a case, the channel appears to be frequency selective, and multipath resolvability is possible. Compared to narrowband CDMA, beyond multipath exploitation, WCDMA presents enhanced performance through certain advantages, such as a decrease of the required transmitted power to achieve a given performance, greater tolerance to power-control errors, fading-effects reduction, the capability to transmit higher data rates and multimedia traffic, and so forth.

Figure 4. Transmitter and receiver models of DS/CDMA

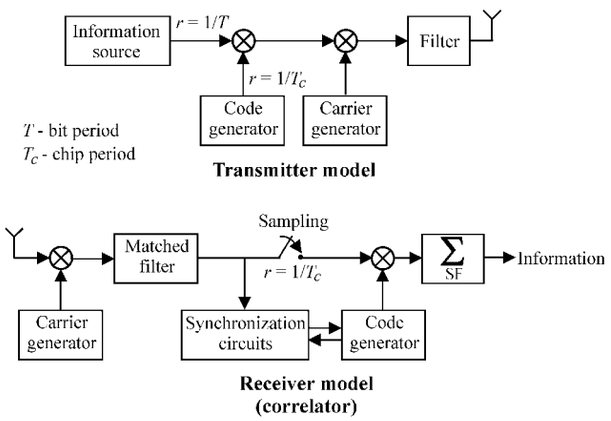
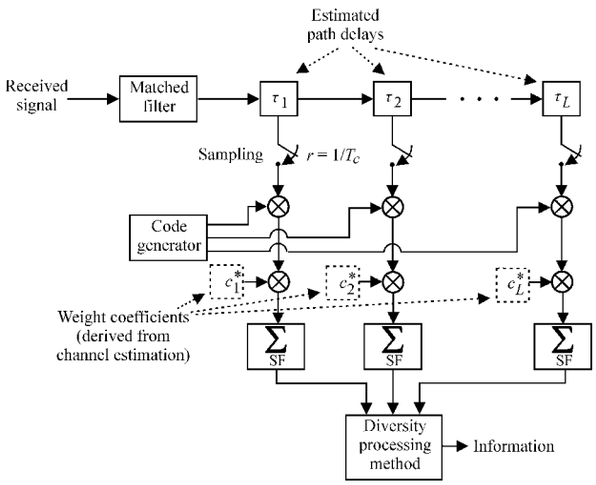


Figure 5. RAKE receiver realization example



CONTEMPORARY APPLICATION OF WCDMA-FDD FOR UMTS

This section summarizes the basic concepts and procedures of applied WCDMA-FDD systems (based on 3GPP [3rd Generation Partnership Project] Rel. 99 or at most Rel. 4).

Information Organization

Source information arrives in transmission time intervals (TTIs) of 10, 20, 40, or 80 ms. Information bits are organized in blocks and CRC (Cyclic Redundancy Check) attachment, forward error-correction (FEC) coding, rate matching, interleaving, and information multiplexing are applied (Holma & Toskala, 2000). FEC can be convolutional of rate 1/2, 1/3, or turbo of rate 1/3, depending on the information type. The produced channel symbols are of rate $7.5*2^m$ (m = 0 to 7) Ksps. Examples of coding and multiplexing are given in 3GPP TR (Technical Report) 25.944.

Multiple-Access Methodology

Multiple access is realized through channelization and scrambling. Channel symbols are spread by the channelization code (orthogonal Hadamard codes) and then chip-by-chip multiplication with the scrambling code takes place (long, partial gold codes of length 38,400 or short S(2) codes of length 256 for future uplink multiuser reception). The chip rate of both channelization and scrambling codes is constant at 3.84 Mchip/s.

Figure 6. Multiple-access methodology of WCDMA-FDD

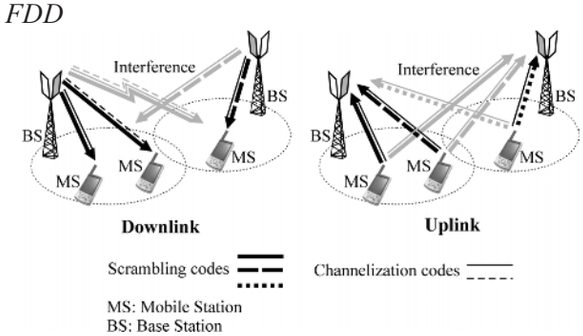


Figure 7. Example of OVSF channelization code-tree usage

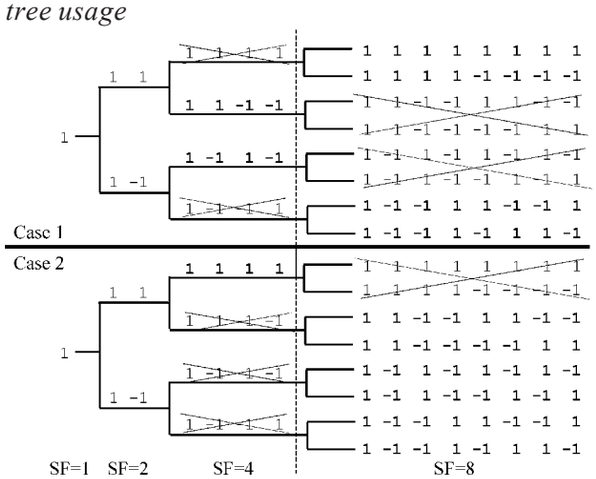


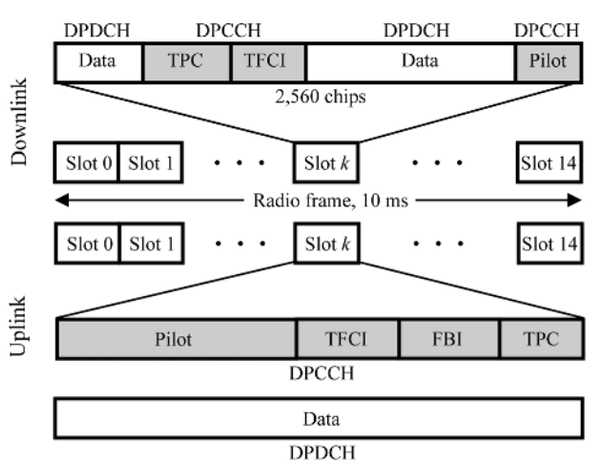
Table 1. Commercial RABs provided with UMTS Rel. 4 (mid-2004)

RAB class	Service	Bit rate (UL/DL) in Kbps	Channel symbol rate (UL/DL) in Ksps *	Spreading factor (UL/DL)
Conversational	CS AMR speech	12.2/12.2	60/30	64/128
	CS data	64/64	240/120	16/32
Streaming	CS streaming	/57.6	/120	/32
Interactive/ background	PS data	64/64	240/120	16/32
		64/128	240/240	16/16
		64/384	240/480	16/8
MultiRAB	CS AMR speech, PS data	12.2/12.2 64/64	240/120	16/32

* Including physical layer control information
AMR: Adaptive multi-rate codec

In the uplink, each user is assigned a unique (among 2^{24} available) scrambling code. Channelization codes are used for separating data and control streams between each other and may have lengths and SFs equal to 2^k (k = 2 to 8) chips. Data-rate variability is achieved by alternating the length of the channelization code (SF) that spreads information symbols. The greater the SF, the lower the information rate is, and vice versa. Parallel usage of more than one channelization code for high uplink data rates (multicode operation) is allowed only when SF equals 4, and this has not been commercially applied yet.

Figure 8. UL/DL dedicated channels of WCDMA-FDD (see also 3GPP TS 25.211)



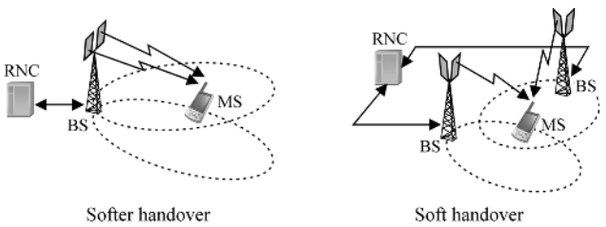
TPC: Power-control command bits (1 command per TS)
TCFI: Transport format combination indicator (Indicates data format)
FBI: Feedback indication bits (used when transmit diversity is engaged)
Pilot bits: Used for coherent reception

Downlink separation is twofold. Cells are distinguished between each other by using different primary scrambling codes (512 available). Each intracell user is assigned uniquely an orthogonal channelization code (SF = 2^k, k = 2 to 9). The same channelization-code set is used in every cell (Figure 6). In order to achieve various information rates (by different SFs) while preserving intracell orthogonality, the channelization codes form an orthogonal variable spreading factor (OVSF) code tree. While codes of equal length are always orthogonal, different length codes are orthogonal under the restriction that the longer code is not a child of the shorter one. Such cases are displayed in Figure 7. Additional scrambling codes (secondary) can be used in the cell mainly for enhancing capacity by reusing the channelization codes. Table 1 summarizes commercially available radio-access bearers (RABs) for circuit-switched (CS) and packet-switched (PS) services.

Transmission

Data transmission is organized in frames of 10 ms (38,400 chips) consisting of 15 TSs (2,560 chips). Each TS contains information data (DPDCH – Dedicated Physical Data Channel) and physical-layer signaling (DPCCH – Dedicated Physical Control Channel;Figure 8). DPDCH and DPCCH are quadrature multiplexed before being scrambled in the uplink and time multiplexed in the downlink. Modulation at the chip level is quadrature phase shift keying (QPSK) in both uplink and downlink. Demodulation is coherent.

Figure 9. Soft and softer handover principle



Soft Handover

Soft handover is the situation when a mobile station communicates with more than one cell in parallel, receiving and transmitting identical information from and to the cells (Figure 9). The cells serving the MS consist of the active set and may belong to the same (softer handover) or other BSs. A combination of cells' signals takes part in the MS RAKE receiver in the downlink, and in the BS RAKE (in the softer case) or in the RNC (radio network controller; signal selection through CRC error counting) in the uplink. Gains in the order of 2 dB (in the signal-to-interference ratio [SIR]) have been reported with soft handover, resulting in enhanced-quality reception. The drawback is the consumption of the limited downlink orthogonal-code resources of the active-set cells.

Power Control

Fast power control (1,500 Hz) is very important for overcoming the near-far problem, reducing power consumption for acceptable communication, and eliminating fading by a significant amount for relatively low-speed moving MSs. Fast power control is based on achieving and preserving a target SIR value (set with respect to information type). An outer power control is used in the uplink for adjusting the SIR target to catch the MSs speed and environmental changes. Power-control-balancing techniques are employed in the downlink to prevent large power drifts between cells during soft handover.

Capacity, Coverage, and Information-Rate Considerations

The capacity and coverage of the system are dynamically adjusted according to the specific conditions and load. Rather involved RRM (radio resource management), admission, and congestion-control strategies have been developed to guarantee acceptable quality of service. In any case, there are some limiting factors that need to be addressed regarding the capacity and information-rate capability of the system.

Speaking of the uplink, 64 Kbps (circuit or packet) is the achievable standard information rate commercially available. Total cell throughputs in the order of 1.5 Mbps have been predicted for microcells (Holma & Toskala, 2000). Additionally, admission-control parameterizations normally assume a reception of 120 to 180 ASEs for uplink cells (air-interface speech equivalent; 1.6 ASEs for voice, 11.1 for 64 Kbps CS, 8.3 for 64 Kbps PS; Ericsson, 2003), that is, about 100 voice users or about 15 users of 64 Kbps at most. Noting that urban 2G (2nd Generation) microcells support more than 80 Erlangs of voice traffic during busy hours, 3G voice capacity seems to be adequate. However, the capacity of higher rate services still seems well below the acceptable limit for mass usage. A rate of 384 Kbps is possible when SF equals 4. Low SF, however, means low tolerance to interference and the need for a power increase. The power capability of MSs may not be enough for achieving the desired SIR when SF equals 4, especially when they are far positioned from the Node-B. Thus, the usage radius of high information rates in a fully developed network would be in the order of decades of meters and for a small number of users (because of the large amount of MAI they produce for coexistent users).

While coverage is uplink limited, capacity is clearly downlink limited (Holma & Toskala). The capacity limitation of the downlink (which should normally support higher rate services) is threefold: the BS power-transmission limitation, limited downlink orthogonality, and the cost limitation of complex MS receivers. The initiation of each new user in the system presupposes enough BS power and OVSF tree-branch availability to support the requested data. Besides this, the initiation of new users poses additional interference to other cell MS receivers. More-

over, the downlink is more sensitive to environmental differences. Although multipath may enhance performance when MSs engage adequate RAKE branches, it also leads to intracell orthogonality loss. Downlink throughputs of the order of 1.5 Mbps have been predicted (Holma & Toskala).

WCDMA-FDD ENHANCEMENTS

Downlink

Transmit Diversity and MIMO Systems

Although not yet commercially applied, transmit diversity methods have been early specified for performance enhancement (space-time transmit diversity [STTD]; 3GPP TS 25.211). Each cell engages two transmit antennas and a proper coding procedure consisting of transmitting identical information in a different order and format (space-time block coding), resulting in extra diversity reception at the receiver. The system may operate in either an open-loop or a closed-loop format, where feedback information (FBI bits) is used to adjust the transmission gains of the antennas. It has been demonstrated (Bjerke, Zvonar, & Proakis, 2004; Vanganuru & Annamalai, 2003) that gains of more than 5 dB (in SNR – Signal to Noise Ratio) can be achieved with a single receiving antenna for open-loop schemes when compared to no transmission diversity. Closed-loop schemes provide an extra 3 dB gain, while engaging a second receiving antenna at the MS enhances performance by 3 to 4 dB more. Transmit diversity is a special case of the MIMO (multiple-input, multiple-output) concept that has gained great interest in the last decade (Molisch & Win, 2004). MIMO systems may be used for diversity enhancement or spatial-information multiplexing for information-rate increase. However, the implementation of multiantenna systems, especially for the MS, is still too costly. MIMO systems will play a significant role in WCDMA enhancement, but their commercial use is still far.

Advanced Receivers

The evolution of MS receivers will yield capacity enhancement of the system. Several strategies have been proposed. The key concept is to minimize orthogonality loss that arises from multipath propagation. The proposed methods employ MMSE (minimum mean squared error) receivers for chip-level equalization (Hooli, Latva-aho, & Juntti, 1999) or generalized RAKE receivers (Bottomley, Ottosson, & Wang, 2000) with tap gains derived from a maximum-likelihood criterion. These schemes produce gains in the order of 2 to 3 dB (uncoded performance) over the standard RAKE structure. Additionally, multipath interference cancellers (MPICs) have been considered (Kawamura, Higuchi, Kishiyama, & Sawahashi, 2002) with comparable performance (slightly worse) and lower complexity that reaches (in high SIRs) flat fading performance. Maximum-likelihood sequence estimation (MLSE) on the chip level is another method that guarantees even better performance at a higher complexity (Tang, Milstein, & Siegel, 2003). In any case, advanced receivers' adoption for commercial use will pose a complexity and performance equilibrium as a selection point for manufacturers and end users (with respect to cost).

HSDPA and Link Adaptation Methods

High-speed downlink packet access (HSDPA; Honkasalo et al., 2002) is a key enhancement over the 2-Mbps packet data of WCDMA. Starting with 3GPP TR 25.848 (V4.0.0), a certain number of changes where finally adopted in Rel. 5 (3GPP TS 25.308 V.5.5.0), a brief description of which is given below.

A new downlink data-traffic physical channel was introduced (HS-PDSCH – High Speed-Physical Downlink Shared Channel) with a frame of 3 TSs (2 ms), called a subframe, and a corresponding TTI, which allows faster changes of transmitted formats. SF is always set to 16. A specific part of the code tree is assigned to HSPDA (5 to 12 codes). Each HS-PDSCH is associated with a DPDCH channel, and a single information transmission to an MS is allowed in each subframe.

Adaptive modulation and coding (AMC) engaging higher order modulation schemes that span the region from QPSK to 64 QAM (quadrature amplitude modulation) regarding modulation, and 1/4 to 3/4 turbo coding, were proposed and evaluated (Nakamura, Awad, & Vagdama, 2002). The idea is to adapt the modulation-coding scheme (MCS) to the changing

channel conditions using reverse-link channel-quality indicators (CQIs) in order to achieve higher throughputs. In Rel. 5 and 6 (3GPP TS 25.308 V.5.5.0, V.6.1.0) QPSK and 16 QAM were finally specified for usage, reaching a bit rate of 720 Kbps. Multicode transmission further improves information rates (Kwan, Chong, & Rinne, 2002), exceeding 2-Mbps throughputs per cell for low-speed terminals with five parallel codes.

Hybrid ARQ (autorepeat request) is engaged for enhancing performance by retransmissions (according to acknowledgement messages by the MS) when packets are received in error. Incremental redundancy (more parity bits at retransmissions) or chase-combining techniques (identical retransmission) may be used.

Fast cell selection is specified, where the users are served by the best cell of the active set at each instance, decreasing, in this way, interference, and increasing capacity.

Fast transmission scheduling is also considered, including the sequential order of serving MSs (round-robin), max C/I (Carrier to Interference) selection (where the best user is served in each TTI), or proportionally fair serving that is a trade-off between throughput maximization and fairness.

The above techniques, in conjunction with MIMO methods, promise rates that reach about 7 Mbps with advanced receiver structures (Kawamura et al., 2002). Some further enhancements under consideration (3GPP TR 25.899 V.6.0.0) include multiple simultaneous transmissions to the MS in the duration of a subframe. In this way, scheduled retransmission can be multiplexed with new transmissions to provide higher throughput. OVSF code sets can be reused in a partial (only for HSDPA codes) or full manner by using a secondary scrambling code in conjunction with two transmit antennas where each one is scheduled to transmit to specific users according to interference experienced in the downlink. Fast, adaptive emphasis for users in a soft handover with closed-loop STTD will exist, where the antenna gains are set according to the existence or nonexistence of downlink HSDPA information. Fractional, dedicated physical channels where the associated dedicated physical channels of different users can be multiplexed together in a single downlink code in order to reduce downlink code-set consumption will be implemented.

Uplink

The main approaches for uplink enhancement, in terms of performance and information rate, are multiuser detectors or interference cancellers, and adaptive multirate techniques.

Increasing the complexity in BS receivers will be inevitable for increasing uplink capacity. Being out of this discussion because of its great complexity, optimal multiuser detection (MUD; Verdu, 1998) gave rise to blind, adaptive MUD approaches that avoid perfect knowledge of all user codes and channel statistics either with training sequences or without. The second approach is based on step-by-step orthogonalization (MMSE) of the desired signal to the interference by adding a proper varying vector. Newer techniques can cope with multipath interference as well as with proper precombining and windowing methods, aiming to orthogonalize the total interference signal received from all RAKE branches (Mucchi, Morosi, Del Re, & Fantacci, 2004). This method (which can be used either in the UL or DL) achieves great performance, approaching single-user behavior with near-far resistance. Other interesting methods are based on interference cancellation in conjunction with beam-forming techniques and space-time combining (Mottier & Brunel, 2003). The idea is to reproduce interference iteratively and cancel it from the desired signal. Near-single-user performance is achieved.

Multicoding has also been proposed for uplink communication. Multicode multidimensional and multicode-OVSF schemes (Kim, 2004) with proper receivers have been evaluated for reliable, high uplink transmission rates. Adaptive schemes with rate adaptation (multiple SFs) have also been analyzed. It was found that the optimum combined rate-power adaptation scheme achieves great performance (Jafar & Goldsmith, 2003), with rate adaptation being the main contribution factor. In this context, Yang and Hanzo (2004) showed that with adaptive SF, a 40% enhancement of total throughput is achieved (single-cell evaluation) without extra power consumption, quality degradation, and interference increase.

Uplink enhancement is already under consideration in 3GPP (TR 25.896 V.6.0.0), which introduces an enhanced uplink-dedicated channel (E-DCH) that is code multiplexed (multicode operation) with a different SF than normal dedicated channels, shorter

frame size (2 ms), uplink H-ARQ (Hybrid ARQ), and so forth. Results show a 50 to 70% cell-throughput enhancement compared to an R99 uplink.

Beam-Forming Techniques

Beam-forming techniques for both the downlink and uplink are based on the use of antenna arrays that focus the antenna beam to a specific region or user in order to improve the SIR. Several techniques have been studied (Li & Liu, 2003) and are under consideration for future use. The capacity enhancement achieved has the drawback of further installation and optimization costs for already operating networks. Thus such techniques will be rather engaged by new operators. It should be mentioned though that cost savings from the usage of common 2G to 3G antennas are lost.

CONCLUSION

The present status of commercial WCDMA and its future trends have been addressed. Entering the multimedia era forces operators to follow WCDMA enhancements as soon as possible in order to keep and expand their subscribers' base. While first-launch implementation costs may not have been yet amortized, a future glance of 4G (4th Generation) and WLAN competition obligates operators to implement new WCDMA techniques and offer new services as soon as customers are ready to follow. Under these circumstances, the most cost-effective solutions will be selected. Among the different technologies described, transmit diversity schemes, HSDPA for downlink and link-adaptation methods and advanced receivers for uplink seem to be the next commercial step since the cost encumbers the operator. These techniques will strengthen the potential of multimedia provision and will provide adequate capacity and quality that will place an advantage of UMTS over 4G alternatives.

REFERENCES

Bjerke, B. A., Zvonar, Z., & Proakis, J. G. (2004). Antenna diversity combining schemes for WCDMA systems in fading multipath channels. *IEEE Transactions on Wireless Communications, 3*(1), 97-106.

Bottomley, G. E., Ottosson, T., & Wang, Y.-P. E. (2000). A generalized RAKE receiver for interference suppression. *IEEE Journal on Selected Areas in Communications, 18*(8), 1536-1545.

Ericsson, A. B. (2003). *Capacity management WCDMA RAN* (User description, 73/1551-HSD 101 02/1 Uen B).

Holma, H., & Toskala, A. (2000). *WCDMA for UMTS*. New York: John Wiley & Sons.

Honkasalo, H., Pehkonen, K., Niemi, M. T., & Leino, A. (2002). WCDMA and WLAN for 3G and beyond. *IEEE Wireless Communications, 9*(2), 14-18.

Hooli, K., Latva-aho, M., & Juntti, M. (1999). Multiple access interference suppression with linear chip equalizers in WCDMA downlink receivers. *Proceedings of GLOBECOM'99*, (pp. 467-471).

Jafar, S. A., & Goldsmith, A. (2003). Adaptive multirate CDMA for uplink throughput maximization. *IEEE Transactions on Wireless Communications, 2*(2), 218-228.

Kawamura, T., Higuchi, K., Kishiyama, Y., & Sawahashi, M. (2002). Comparison between multipath interference canceller and chip equalizer in HSDPA in multipath channel. *Proceedings of VTC 2002*, (pp. 459-463).

Kim, D. I. (2004). Analysis of hybrid multicode/variable spreading factor DS-CDMA system with two-stage group-detection. *IEEE Transactions on Vehicular Technology, 53*(3), 611-620

Kwan, R., Chong, P., & Rinne, M. (2002). Analysis of adaptive modulation and coding algorithm with the multicode transmission. *Proceedings of VTC 2002*, (pp. 2007-2011).

Li, H.-J., & Liu, T.-Y. (2003). Comparison of beamforming techniques for W-CDMA communication systems. *IEEE Transactions on Vehicular Technology, 52*(4), 752-760.

Milstein, L. B. (2000). Wideband code division multiple access. *IEEE Journal on Selected Areas in Communications, 18*(8), 1344-1354.

Molisch, A., & Win, M. Z. (2004). MIMO systems with antenna selection. *IEEE Microwave Magazine, 5*(1), 46-56.

Mottier, D., & Brunel., L. (2003). Iterative space-time soft interference cancellation for UMTS-FDD uplink. *IEEE Transactions on Vehicular Technology, 52*(4), 919-930.

Mucchi, L., Morosi, S., Del Re, E., & Fantacci, R. (2004). A new algorithm for blind adaptive multiuser detection in frequency selective multipath fading channel. *IEEE Transactions on Wireless Communications, 3*(1), 235-247.

Nakamura, M., Awad, Y., & Vagdama, S. (2002). Adaptive control of link adaptation for high speed downlink packet access (HSDPA) in W-CDMA. *Proceedings of Wireless Personal Multimedia Communications Conference*, (pp. 382-386).

Simoens, S., Pellati, P., Gosteau, J., Gosse, K., & Ware, C. (2003). The evolution of 5 GHz WLAN toward higher throughputs. *IEEE Wireless Communications, 10*(6), 6-13.

Tang, K., Milstein, L. B., & Siegel, P. H. (2003). MLSE receiver for direct-sequence spread-spectrum systems on a multipath fading channel. *IEEE Transactions on Communications, 51*(7), 1173-1184.

Vanganuru, K., & Annamalai, A. (2003). Combined transmit and receive antenna diversity for WCDMA in multipath fading channels. *IEEE Communications Letters, 7*(8), 352-354.

Verdu, S. (1998). *Multiuser detection*. Cambridge, UK: Cambridge University Press.

Viterbi, A. J. (1995). *Principles of spread spectrum communication*. Reading, MA: Addison-Wesley.

Yang, L.-L., & Hanzo, L. (2004). Adaptive rate DS-CDMA systems using variable spreading factors. *IEEE Transactions on Vehicular Technology, 53*(1), 72-81.

KEY TERMS

Admission Control: The algorithms used by the system to accept or refuse new radio links (e.g., new users) in the system.

BS: Base station, also referred to as Node-B in UMTS.

Coherence Bandwidth: The bandwidth over which the channel affects transmitted signals in the same way.

Congestion Control: The algorithms used to detect and solve system-overload situations.

CRC (Cyclic Redundancy Check): Block codes used for error detection.

Cross-Correlation: The sum of the chip-by-chip products of two different sequences (codes). A measure of the similarity and interference between the sequences (or their delayed replicas). Orthogonal codes have zero cross-correlation when synchronized.

MLSE: Maximum-likelihood sequence estimation

MMSE: Minimum mean squared error

Multipath Propagation: The situation where the transmitted signal reaches the receiver through multiple electromagnetic waves (paths) scattered at various surfaces or objects.

Near-Far Effect: The situation where the received power difference between two CDMA users is so great that discrimination of the low-power user is impossible even with low cross-correlation between the codes.

RNC (Radio Network Controller): The network element that manages the radio part of UMTS controlling several Node-Bs.

RRM (Radio Resource Management): The algorithms used by the system (RNC) to distribute the radio resources (such as codes or power in UMTS) among the users.

WCDMA-FDD (Wideband Code-Division Multiple Access, Frequency-Division Duplex): The variant of UMTS WCDMA where UL and DL communication are realized in different frequency bands. In the TDD (time-division duplex) variant, UL and DL are realized in different time slots of the frame. TDD has not been applied commercially yet.

Fiber to the Premises

Mahesh S. Raisinghani
Texas Woman's University, USA

Hassan Ghanem
Verizon, USA

INTRODUCTION

Subscribers had never thought of cable operators as providers of voice services, or telephone companies as providers of television and entertainment services. However, the strategies of multiple system operators (MSOs) and telecommunication companies (telcos) are changing, and they are expanding their services into each other's territory. The competition between the MSOs and the telcos is just brewing up.

Many factors influence communications carriers' future and strategies. Among these factors are Internet growth, new Internet Protocol (IP) services such as Voice over IP (VoIP), regulatory factors and strong competition between the carriers. In the past, RBOC's have centered their competition among each other and ignored the threat of the cable MSOs. The cable modem service has a bigger market share than the digital subscriber line (DSL) service, and as the concept of the VoIP technology is being refined and validated, the cable companies will become major players in providing this service at a cheaper price than the regular telephone service and will compete with the RBOCs. Incumbent carriers are seeking ways to encounter the cable MSOs' threat.

BACKGROUND

RBOCs are concerned about the VoIP technology, since this concept will pose a serious threat to their voice market. Vonage, a leader in VoIP over Broadband (VoB), has about 50,000 subscribers, compared to 187.5 million access lines that the RBOCs have. Cable operators can move into the telcos' territory and offer VoB as they did with Internet access. The cable companies could do this by offering this service through a partnership or by building their own services.

The VoB service is offered to broadband subscribers whether they are cable modem or DSL users. VoB providers do not have their own networks; they simply use the cable MSOs' or the telcos' broadband networks to carry their services. The appeal of the VoB services is the result of its cheaper packages. VoB companies such as Vonage and Packet8 are targeting cable MSOs as partners. For cable companies, this would create a bundle that includes cable modem services and VoB, which will provide a great appeal to the subscriber. Cable MSOs already are in the lead in providing broadband services to subscribers; by adding VoIP via broadband, they will be able to offer telephony at lower prices and have another advantage over the telcos.

Major cable operators have announced their interest in VoIP technology. Time Warner Cable has formed an alliance with MCI and Sprint, and the group has announced that by the end of 2004 it will offer VoIP to 18 million subscribers. Comcast is another cable operator already in the process of testing VoIP in many states, and will offer this service in the nation's largest 100 cities (Perrin et al., 2003a). The MSOs have continued to upgrade their networks to have a bigger share of Internet access and to enter the lucrative voice market. On the other hand, the telcos have continued to develop their networks around DSL and voice service, ignoring television and video services (Jopling & Winogradoff, 2002).

FIBER TO THE PREMISES (FTTP)

To deal with the threat of VoB providers, telcos have to upgrade their networks to compete with the cable

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

MSOs. FTTP is a potential alternative to DSL. It is a great initiative to meet the growing demand of consumers and business to a faster Internet connection and reliable medium for other multimedia services. Since signals will travel through fiber optic networks at the speed of light, FTTP delivers 100 mega bits per second (Mbps), as opposed to 1.5 Mbps for DSL. Thus, FTTP delivers a higher bandwidth at a lower cost per megabyte than alternative solutions. This substantially increased speed will enable service providers to deliver data, voice and video ("triple play") to residential and business customers. As a result of this increase in speed, a new breed of applications will emerge and open horizons for the RBOCs to venture into a new territory. The deployment of FTTP will help eliminate the bandwidth limitations of DSL. DSL will still be a key player for the near future, but in the long run, DSL customers will be migrated to the new fiber network. FTTP will pave the way for the RBOCs to compete head to head with cable providers. Comcast Corp., based in Philadelphia, is the largest cable provider based in the United States. It is upgrading some of its customers' Internet services to 3 mega bits per second, which is significantly more than what phone companies can offer through their DSL network. FTTP will simulate competition in the communication industry and entertainment providers, and will provide RBOCs a medium with which to compete against cable companies.

FTTP COMMON SPECIFICATIONS AND EQUIPMENT

In May 2003, BellSouth, SBC and Verizon agreed on common specifications for FTTP. This agreement has paved the way for suppliers to build one type of equipment based on the specifications provided by the three companies. By mid September, the three companies had short-listed the suppliers, and the equipment was brought to labs to be tested by the three companies, where they will select finalists based on the test results and proposals. The technology being evaluated is based on the G.983 standard for passive optical network (PON) (Hackler, 2003). This standard was chosen based on its flexibility to support Asynchronous Transfer Mode (ATM) and its capacity to be upgraded in the future to support either ATM or Ethernet framing.

As the cost of electronic equipment has fallen dramatically in recent years, it is more feasible now to roll out FTTP than it was a few years ago. Many equipment manufacturers, such as Alcatel, Lucent, Nortel and Marconi, are trying to gain contracts from the big three RBOCs to manufacture and provide FTTP components. The bidding war for these contracts will be very competitive, and providers have to choose equipment suppliers based on the price and specifications of the equipment.

REGULATORY ENVIRONMENT AND THE FCC ORDER

The regulatory environment will also be a major factor in the progress of the FTTP rollout. At the time of this writing, it was still unclear how the Federal Communications Commission (FCC) will handle this issue. Service providers are optimistic that the FCC decisions will favor them. RBOCs are hoping that the FCC will provide a clear ruling regarding national broadband networks.

WHY INVEST IN FTTP AND NOT UPGRADE COPPER?

Several existing technologies can accommodate the triple-play services. For example, Asynchronous DSL (ADSL) is a broadband technology that can reach 8-10 Mbps, and ADSL2 has an even higher range of 20 Mbps. The ADSL technology can be deployed with a fast pace by using existing copper wiring. The disadvantage of the copper-based networks and DSL technology is that they have a regulatory constraint to be shared with competitors, which makes it less attractive to invest in this medium. Another disadvantage is that signals do not travel a long distance. They need expensive electronic equipment to propel the signal through. This expensive equipment will result in high maintenance and replacement costs. Another weakness of DSL technology is that the connection is faster for receiving data than it is for sending data over the Internet.

Another broadband technology to be considered is cable technology. It has bandwidth capacity in the range of 500 Kbps to 10 Mbps. It can deliver data, voice and video and is 10 times faster than the telephone line. But cable technology has its weaknesses, too. It is less reliable than DSL and has a limited upstream bandwidth, which is a significant problem in peer-to-peer applications and local Web servers. Another weakness is that the number of users inversely impacts performance and speed of the network (Metallo, 2003).

On the other hand, fiber will deliver a higher bandwidth than DSL and cable technologies, and maintenance costs will be lower in comparison with copper-based networks. As mentioned earlier, FTTP will use PON, which will minimize the electronic equipment needed to propel the signals. Once the network is in place, the cost of operations and maintenance will be reduced by 25%, compared with copper-based networks.

One company that already had a head start with FTTP technology is Verizon. It will invest $2 billion over the next 2 years to roll out the new fiber network and replace traditional switches with softswitches (software switches). These softswitches will increase the efficiency of the network and eliminate wasted bandwidth. The traditional circuit switch architecture establishes a dedicated connection for each call, resulting in one channel of bandwidth to be dedicated to the call as long as the connection is established. The new architecture will break the voice into packets that will travel by the shortest way over the new network; as soon as the packet reaches its destination, the connection is broken and no bandwidth is wasted.

At the design level, softswitches and hard switches differ drastically. Features can be added or modified easily in the softswitch, while they have to be built into the hard switch. Also, FTTP will be the means to deliver the next generation of products and services at a faster speed and with more data capacity (i.e., send up to 622 Mbps and receive 155 Mbps of data compared to 1.5 Mbps that DSL or cable modems are capable of (Perrin et al., 2003b). This will create an opportunity to develop and sell new products and services that can only be feasible over this kind of technology, which will result in generating a new revenue stream.

CHALLENGES AND ISSUES WITH FTTP

Fiber will be the access medium for the next 100 years or so. But in the meantime, the migration from copper to fiber must not be viewed as a short-term initiative. It will take 5 to 10 years to become a reality. FTTP technology deployment is very different from DSL deployment. While DSL deployment was an add-on technology, where the network and operations systems impact was relatively minor, FTTP deployment will undertake a new infrastructure and will require major changes in support systems.

Another issue that will be facing RBOCs is DSL technology. RBOCs will be burdened to support DSL technology as well as FTTP technology. By supporting DSL after the complete deployment of FTTP, incumbents will endure extra costs that can be avoided by switching their customers to the FTTP network and abandoning the DSL network.

The FCC Triennial Review Order will put the new technology in jeopardy in case of a negative clarification or ruling. The Telecommunications Act of 1996 requires incumbents to lease their networks to competitors at rates below cost. If copper networks are retired, incumbents are required to keep providing competitors with a voice-grade channel. Another challenge will stem from providing video to customers. Cable companies are well established in this domain and have the advantage over RBOCs. In the 1980's and 1990's, RBOCs attempted to expand into entertainment services and failed miserably. But, their new strategies to enter entertainment services have to be well planed, and the companies have to learn form their previous failures.

FTTP DEPLOYMENT

Networks are laid down via aerial and carried via poles or cables under the ground. Using this infrastructure will enable FTTP to be deployed at a relatively low cost. The fiber will be installed at close proximity to the customer and will be extended to new build areas when needed. When the customer requests the service, rewiring will be added from the Optical Network Terminals (ONT) to the

customer's premises. This will keep a close correlation factor between deployment costs and return on capital. This will tie some of the FTTP expenditure to customer demand. Deployment can be started in areas where the highest revenue is generated, and move to other areas where the infrastructure is the oldest.

FUTURE OUTLOOK

FTTP is promising to deliver the next-generation network with an advanced bandwidth. It will also be capable of delivering various services that may be available in the next few years. In 2003, Microsoft showed a beta demonstration of a live online gaming application with simultaneous voice, video and data services over FTTP. A new breed of gaming applications will be feasible with FTTP. Applications such as peer-to peer, where music files, video files and large data files are exchanged, would become more attractive with FTTP. Another application that will become viable is video on demand (VOD), where customers can order any movie of their choice at any time. Beardsley (2003) reports that broadband offers a new distribution path for video-based entertainment; a medium for new interactive-entertainment services (such as interactive TV) that need a lot of bandwidth; and a way to integrate several media over a single connection. For example, FastWeb, based in Milan, Italy, can now supply 100,000 paying households in Italy with true VOD, high-speed data and digital voice, all delivered over a single optical-fiber connection. As mature markets reach scale with large online audiences, broadband may start to realize some of its underlying—and long-hyped—potential for advertisers. The new advanced broadband will become a platform for many industries to deliver marketing, sales and communications services. Broadband is already changing the way companies do business and could alter the way markets work. For instance, remote learning will improve greatly, and educational institutions will be in a better position to offer services in remote locations. Many other fields, such as health care, public sector, retail and financial services, will also see the positive impact of broadband.

CONCLUSION

The incumbent carriers have to encounter cable operators' attacks on their telephony market by expanding their services and offering similar services that the cable companies offer and more. Triple play will give the telcos an advantage over cable operators, or at least an equal edge. They have to deploy broadband networks that can provide subscribers with entertainment/television services, to compensate the telcos' loss of revenues incurred from cable operators' deployment of VoIP. For telcos, expanding into the entertainment market has to be based on the strategy of meeting consumers' changing needs. Cable operators have already raged a battle against satellite companies, who are competing for the same customers. Another factor to be considered is the MSOs' slow adoption of digital technology. On the financial side, MSOs are weaker than telcos.

The RBOCs are faced with staggering sinking costs. The idea is to come up with new killer applications suited for the new network that will appeal to (potential) customers' tastes and are affordable, to lure subscribers. FTTP has to do more than current systems. Enhancing what the current technology is capable of doing does not justify the cost and/or effort that have to be invested in FTTP. A newer breed of "killer application/s" that would lure the subscriber has to be implemented and delivered with the service. They have to be flexible to tailor their bundle according to the customers' demands and needs.

REFERENCES

Beardsley, S., Doman, A., & Edin, P. (2003). Making sense of broadband. Retrieved March 28, 2004, from *www.mckinsey quarterly. com/ article_ page.asp?ar=1296&L2=38&L3=98*

Cherry, S.M. (2004). Fiber to the home. Retrieved February 8, 2004, from *www.spectrum.ieee.org/ WEBONLY/publicfeature/jan04/0104comm3.html*

Federal Communications Commission (2004). *FCC to consider VoIP regulation.* Retrieved March 28, 2004, from *http://eweb.verizon.com/news/vz/ 010104/story11.shtml*

Hackler, K., Mazur, J., & Pultz, J. (2003). Incumbent carriers link up to cut fiber cost. Retrieved February 7, 2004, from *www4.gartner.com/DisplayDocument?id=396501&ref=g_search#h1*

Jopling, E., & Winogradoff E. (2002). Telecom companies, cable operators battle for consumers. Retrieved March 28, 2004, from *www3.gartner.com/resources/111900/111916/111916.pdf*

Metallo, R. (2003). As Fiber-to-the-Premises enters the ring …. Retrieved March 10, 2004, from *www.lucent.com/livelink/090094038 0059ffa_White_paper.pdf*

Perrin, S., Harris, A., Winther, M., Posey, M., Munroe, C., & Stofega, W. (2003a, July). *The RBOC FTTP initiative: Road map to the future or déjà vu all over again.* Retrieved February 8, 2004, from *www.idc.com/getdoc.jhtml?containerId=29734*

Perrin, S., Stofega, W., & Valovic, T.S. (2003b, Sept). *Voice over broadband: Does Vonage have the RBOCs' number?* Retrieved February 23, 2004, from *www.idc.com/getdoc.jsp?containerId=30020&page*

White, J. (2003). *Verizon Communications – Taking fiber to the subscriber.* Retrieved February 25, 2004, from *www.opticalkeyhole.com/keyhole/html/verizon.asp?bhcd2=1079731603*

KEY TERMS

Asynchronous Digital Subscriber Line (ADSL): A digital switched technology that provides very high data transmission speeds over telephone system wires. The speed of the transmission is asynchronous, meaning that the transmission speeds for uploading and downloading data are different. For example, upstream transmissions may vary from 16 Kbps to 640 Kbps and downstream rates may vary from 1.5Mbps to 9Mbps. Within a given implementation, the upstream and downstream speeds remain constant.

Asynchronous Transfer Mode (ATM): A high-speed transmission protocol in which data blocks are broken into cells that are transmitted individually and possibly via different routes in a manner similar to packet-switching technology.

Bandwidth: The difference between the minimum and the maximum frequencies allowed. Bandwidth is a measure of the amount of data that can be transmitted per unit of time. The greater the bandwidth, the higher the possible data transmission rate.

Digital Subscriber Line (DSL): A switched telephone service that provides high data rates, typically more than 1 Mbps.

Fiber Optic Cable: A transmission medium that provides high data rates and low errors. Glass or plastic fibers are woven together to form the core of the cable. The core is surrounded by a glass or plastic layer, called the cladding. The cladding is covered with plastic or other material for protection. The cable requires a light source, most commonly laser or light-emitting diodes.

Internet Protocol (IP): The network layer protocol used on the Internet and many private networks. Different versions of IP include IPv4, IPv6 and IPng (*next generation*).

Multiple Service Operators (MSOs): Synonymous with cable provider. A cable company that operates more than one TV cable system.

Regional Bell Operating Company (RBOC): One of the seven Bell operating companies formed during the divestiture of AT&T. An RBOC is responsible for local telephone services within a region of the United States.

Voice Over IP (VoIP): This is the practice of using an Internet connection to pass voice data using IP instead of the standard public switched telephone network. This can avoid long-distance telephone charges, as the only connection is through the Internet.

Fiber-to-the-Home Technologies and Standards

Andjelka Kelic
Massachusetts Institute of Technology, USA

INTRODUCTION

Fiber-to-the-home (FTTH) refers to the provisioning of narrowband and broadband services to the residential customer over an optical cable rather than traditional copper wiring. Early trials in the United States, England, and France to provide telephone and broadcast video service to residential customers occurred in the mid- to late 1980s, however, widespread deployment did not follow from these trials (Esty, 1987; Rowbotham, 1989; Shumate, 1989; Veyres & Mauro, 1988). Studies conducted at the time suggested that consumer demand for video and telephone service was not sufficient to warrant the funds necessary for wide-scale deployment of the systems (Bergen, 1986; Sirbu & Reed, 1988).

The studies did not foresee the interest in residential broadband service spurred by the growth of the commercial Internet and the World Wide Web. Since the days of the early trials, residential and small-business lines providing at least symmetric 200-kbps services have grown to 18.1 million as of December 2003 in the United States alone (Federal Communications Commission, 2004), and FTTH has been standardized with an eye toward providing multimedia services.

BACKGROUND

Deployment of residential broadband has been growing around the world. The most commonly deployed technologies are DSL (digital subscriber line) and cable modems (Ismail & Wu, 2003). Wireless for residential broadband also has a small showing.

Both DSL and cable modem services run over existing copper or hybrid fiber-copper plants. The newest DSL technology is VDSL (very-high-fatarate digital subscriber line), which promises to deliver asymmetric speeds of up to 52 Mbps from the provider to the customer (downstream) and 6 Mbps from the customer to the provider (upstream), or symmetric speeds of 26 Mbps (The International Engineering Consortium, n.d.). Unfortunately the technology is distance limited and the maximum speeds can only be achieved up to a distance of 300 m. Longer distances result in a reduction in speed.

Cable modem services' newest standard, DOCSIS 2.0 (Cable Television Laboratories, Inc., 2004), is capable of a raw data rate of 40 Mbps in the downstream and 30 Mbps in the upstream. However, due to the broadcast nature of the system, this bandwidth is typically shared among a neighborhood of subscribers.

Fixed wireless services are also targeting the residential broadband market with a technology capable of up to symmetrical 134.4 Mbps depending on the width of the channel and the modulation scheme used. The technology is known as WiMax and is defined in IEEE 802.16 (Institute of Electrical and Electronics Engineers (IEEE); IEEE, 2002). The original WiMax standard, and the 134.4 Mbps transmission capability, is for use in a frequency range that requires line of sight for transmission. The standard has since been updated via IEEE 802.16a (IEEE, 2003) for use in frequency bands that do not require line of sight for transmission. The drawback to using non-line-of-sight frequency bands is a lower data rate of up to 75 Mbps depending on channel width and modulation scheme. Similar to cable modem service, WiMax also shares its bandwidth among groups of customers.

The technologies under development for fiber to the home promise far greater dedicated bandwidth than any of the proposed future modifications to DSL, DOCSIS, or fixed wireless, and in the case of DSL, over much longer distances. This makes FTTH better suited as a platform to support multimedia services to residential customers.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

FIBER-TO-THE-HOME TECHNOLOGIES AND STANDARDS

FTTH technologies fall into two categories: active or passive. Both types of technologies are capable of delivering voice, video, and data service. Active technologies have an active component such as a switch or router between the central office and the customer. Passive technologies have a passive (unpowered) component, such as an optical splitter, between the central office and the customer.

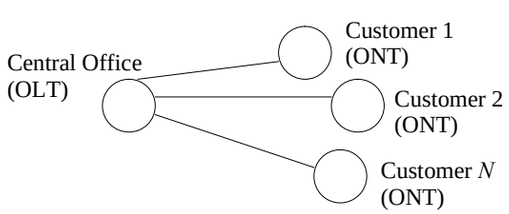
Standards work for FTTH technologies has been taking place in two different organizations: the Institute of Electrical and Electronics Engineers and the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T). The IEEE standards work is focused on the use of Ethernet-based technologies in the access network (Ethernet in the First Mile or EFM) and the ITU standards work (called recommendations) focuses primarily on passive optical networks (PONs). The ITU-T and IEEE standards groups communicate regularly in order to ensure that the standards that are developed do not conflict.

FTTH technologies can be deployed in three different topologies: home run, active star, or passive star (Committee on Broadband Last Mile Technology, National Research Council, 2002).

Home Run

A home-run network topology is a point-to-point topology with a run of fiber from the provider's central-office optical line terminal (OLT) out to each customer optical network terminal (ONT). The fiber run can be either one fiber, with different wavelengths for upstream and downstream transmission, or two separate fibers, one for upstream and one for downstream transmission. A home-run network topology is shown in Figure 1. This architecture is costly because it requires a dedicated fiber for each customer from the central office to the customer premise. The central-office equipment is the only resource that is shared amongst the customer base.

Figure 1. Home-run topology



ITU-T G.985, approved in March 2003, is defined as operating over a point-to-point network topology. G.985 came out of efforts by the Telecommunications Technology Committee (TTC) in Japan to achieve interoperability between vendors for deployed Ethernet-based FTTH systems (ITU-T, 2003c) and has contributed to the EFM Fiber standards work.

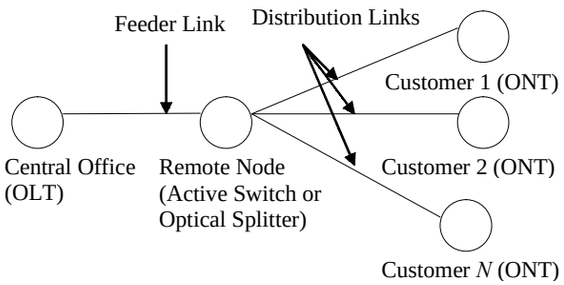
The recommendation describes a single-fiber, 100-Mbps point-to-point Ethernet optical access system. Included are specifications for the optical distribution network and the physical layer, and also the requirements for operation, administration, and maintenance. Transmission is on a single fiber using wave-division multiplexing (WDM), with downstream transmission in the 1480- to 1580-nm range and upstream transmission in the 1260- to 1360-nm range. WDM divides the fiber by wavelength into two or more channels. The standard currently defines a 7.3-km transmission distance with 20- and 30-km distances for further study.

Active Star

In this topology, a remote node with active electronics is deployed between the central office and the customer premises, as shown in Figure 2. The link between the central office and remote node is called the feeder link, and the links between the remote nodes and the customer premises are called distribution links. A star topology is considered more cost effective than a home-run topology because more of the network resources are shared amongst the customers.

EFM Fiber (IEEE 802.3ah) is most commonly deployed in an active star configuration. It is similar in architecture to traditional hubs and switches that run 10BaseF and 100BaseFX today. The standards for EFM Fiber were developed by the IEEE 802.3ah Task Force.

Figure 2. Star topology



The technology consists of point-to-point, single-mode fiber with a range of at least 10 km between the active switch and the ONT. EFM Fiber employs Ethernet and active equipment at speeds of 100 Mbps and 1 Gbps (IEEE 802.3ah Ethernet in the First Mile Task Force, 2004). Operation can be over a single fiber or one fiber for upstream transmission and a second fiber for downstream transmission.

For the two-fiber configuration, transmission is in the 1260- to 1360-nm wavelength band. For operation over a single fiber, upstream transmission is in the 1260- to 1360-nm wavelength band and the downstream transmission wavelength varies depending on the transmission speed. 100-Mbps downstream operation uses the 1480- to 1580-nm wavelength band, and 1-Gbps downstream operation uses the 1480- to 1500-nm wavelength band. The wavelength assignments for 1-Gbps service allow the system to incorporate a dedicated wavelength for broadcast video service in the 1550- to 1560-nm bands as specified by the newer ITU-T passive-optical-network standards described in the following section.

Passive Star

Passive optical networks, or passive star topologies, have no active components between the provider's central office and the subscriber. The remote node of Figure 2 contains an optical splitter in a passive star topology. PONs are point-to-multipoint systems with all downstream traffic broadcast to all ONTs. The PONs under development are ATM-based PONs (asynchronous transfer mode; APONs), gigabit-capable PONs (GPONs), and Ethernet-based PONs (EPONs).

ATM PON

APON systems are PONs that are based on the asynchronous transfer mode. APONs are also known by the name of BPON, or broadband PON, to avoid confusing some users who believed that APONs could only provide ATM services to end users. APONs are defined by the ITU-T G.983 series of recommendations.

ATM uses 53-byte cells (5 bytes of header and 48 bytes of payload). Because of the fixed cell size, ATM implementations can enforce quality-of-service guarantees, for example, bandwidth allocation, delay guarantees, and so forth. ATM was designed to support both voice and data payloads, so it is well suited to FTTH applications.

The APON protocol operates differently in the downstream and upstream directions. All downstream receivers receive all cells and discard those not intended for them based on ATM addressing information. Due to the broadcast nature of the PON, downstream user data is churned, or scrambled, using a churn key generated by the ONT to provide a low level of protection for downstream user data.

In the upstream direction, transmission is regulated with a time-division multiple access (TDMA) system. Transmitters are told when to transmit by receipt of grant messages. Upstream APON modifies ATM and uses 56-byte ATM cells, with the additional 3 bytes of header being used for guard time, preamble bits, and a delimiter before the start of the actual 53-byte ATM cell.

The G.983 series of recommendations define the nominal bit rates for APON to be symmetric 155.52 Mbps or 622.08 Mbps, or asymmetric 622.08 Mbps in the downstream direction and 155.52 Mbps in the upstream direction. The OLT for an APON deployment can support multiple APONs with a split ratio of 32 or 64 subscribers each, depending on the vendor.

ITU-T G.983.1, approved in October 1998, can be deployed as two fibers to each customer (one upstream and one downstream), or, using WDM, as one fiber to each customer. For two fibers, transmission is in the 1260- to 1360-nm band in both upstream and downstream directions. In a single-fiber system, upstream transmission remains in the 1260- to 1360-nm band and downstream transmission is in the 1480- to 1580-nm wavelength band (ITU-T, 1998).

ITU-T G.983.3, approved in March 2001, redefines the downstream transmission band for single-fiber APONs. This allows part of the spectrum to be allocated for video broadcast services or data services. Services can be either bidirectional or unidirectional (ITU-T, 2001).

The wavelength allocations leave the PON upstream wavelengths unchanged at 1260 to 1360 nm. The downstream transmission band is reduced to only include the portion of the band from 1480 to 1500 nm, called the basic band. The enhancement band (Option 1), the 1539- to 1565-nm band, is for the use of additional digital services. The recommendation defines the 1550- to 1560-nm band as the enhancement band (Option 2) for video-distribution service. Two bands are reserved for future use: the band from 1360 to 1480 nm, which includes guard bands, and a future band in the 1480- to 1580-nm range for further study and allocation.

Gigabit PON

Efforts to standardize PON networks operating at above 1 Gbps were initiated in 2001 as the ITU-T G.984 series of recommendations. GPON is a more generalized version of APON and is not dependent on ATM. GPON realizes greater efficiency over APON by not requiring large IP (Internet protocol) packets to be broken up into 53-byte ATM cells. GPON attempts to preserve as many characteristics of the G.983 series of recommendations as possible, however, due to technical issues relating to providing the higher line rates, the two systems are not interoperable (ITU-T, 2004).

As with APON, the system may be either a one- or two-fiber system. In the downstream direction, GPON is also a broadcast protocol with all ONTs receiving all frames and discarding those not intended for them. Upstream transmission is via TDMA and is controlled by an upstream bandwidth map that is sent as part of the downstream frame. GPON uses encryption for the payload. The encryption system used assumes that privileged information, like the security keys to decode the payloads, can be passed upstream in the clear due to the directionality of the PON (i.e., that any ONT in the PON cannot observe the upstream traffic from any other ONT in the PON).

The GPON OLT can support split ratios of 16, 32, or 64 users per fiber with current technology. ITU-T (2003b) G.984.2 anticipates future ratios of up to 128 users per fiber and accounts for this in the transmission-convergence layer. As with G.983.3, for a single-fiber system, the operating wavelength is in the 1480- to 1500-nm band in the downstream and in the 1260- to 1360-nm band in the upstream. This leaves the 1550- to 1560-nm band free for video services. For a two-fiber system, the operating wavelength is in the 1260- to 1360-nm band in both the downstream and the upstream directions.

GPON has seven transmission-speed combinations (line rates): symmetric 1.2 or 2.4 Gbps; or asymmetric 1.2 or 2.4 Gbps downstream with 155 Mbps, 622 Mbps, or 1.2 Gbps in the upstream (ITU-T, 2003a). The physical reach of the GPON is 10 km for speeds of 1.2 Gbps and below, and 20 km for speeds above 1.2 Gbps.

Ethernet PON

EPON is Ethernet over a passive optical network. Similar to EFM Fiber, standards are being developed in the IEEE 802.3ah Task Force. The protocol used in EPON is an extension of Ethernet (IEEE 802.3) and operates at 1 Gbps with a range of 10 or 20 km between the central office and the customer. The architecture is a single shared fiber with an optical splitter, as with other PON architectures. The supported split ratio is 16 users per PON. The system operates in the 1480- to 1500-nm band in the downstream direction, and in the 1260- to 1360-nm band in the upstream direction. As with 1-Gbps EFM Fiber, while not specifically mentioning a wavelength for broadcast video service, EPON allocates its wavelengths to leave the 1550- to 1560-nm band open and is capable of supporting a broadcast video wavelength in that band.

Since Ethernet does not utilize a point-to-multipoint topology, EPON required the development of a control protocol to make the point-to-multipoint topology appear as a point-to-point topology. This protocol is called the multipoint control protocol (MPCP).

Like all PONs, in the downstream direction EPON is a broadcast protocol. Every ONT receives all packets, extracts the Ethernet frames intended for that customer, and discards the rest. As with APON and GPON, transmission in the upstream direction is regulated by TDMA.

Table 1. FTTH single-fiber system summary

Technology	1550-nm Video	Max. Speed (Mbps)	Homes per Feeder	Standard	Year
G.985	No	100	N/A	G.985	2003
EFM Fiber	No	100	N/A	802.3ah	2004
	Yes	1,000			
APON	No	622	16, 32	G.983.1	1998
	Yes			G.983.3	2001
GPON	Yes	2,400	64, 128	G.984	2003
EPON	Yes	1,000	16	802.3ah	2004

FUTURE TRENDS

As shown in Table 1, FTTH standards are moving toward higher line speeds, more users per PON, and standardized wavelengths with the ability to provide a dedicated wavelength for broadcast video service. The GPON recommendations anticipate some of these trends by allowing for wavelengths for future expansion, and the possibility of higher split ratios and line speeds in the formulation of the standard.

The standards for EFM Fiber and G.985 do not specify the number of homes that must be supported per feeder fiber. This allows the systems to be deployed in either an active star or home-run topology supporting as many users as current switching technology is capable of without the need to modify the standard. In some current active star implementations, the number of homes per feeder fiber supported is as high as 48. This number is expected to increase as switching technology improves.

CONCLUSION

The ITU and IEEE are working to develop FTTH standards that do not conflict with one another. These standards are converging toward standardized wavelength allocations for upstream and downstream transmission with the ability to support a consistent, dedicated wavelength for broadcast video service. The standards are also moving toward higher line speeds and the ability to support more users.

Fiber to the home provides greater bandwidth than any of the residential networking alternatives. With the addition of an entire 1-GHz wavelength for broadcast video in the standards for EFM Fiber, EPON, G.983.3 APON, and GPON, FTTH can support HDTV (high-definition television) channels and video-on-demand functions without competing with voice or data bandwidth, making it well suited for multimedia applications.

REFERENCES

Bergen, R. S., Jr. (1986). Economic analysis of fiber versus alternative media. *IEEE Journal on Selected Areas in Communications, 4*, 1523-1526. New York: IEEE.

Cable Television Laboratories, Inc. (2004). *Data-over-cable service interface specifications DOCSIS 2.0: Radio frequency interface specification*. Louisville, CO: Cable Television Laboratories, Inc. Retrieved July 7, 2004, from *http://www.cablemodem.com/downloads/specs/SP-RFIv2.0-I05-040407.pdf*

Committee on Broadband Last Mile Technology, National Research Council. (2002). *Broadband: Bringing home the bits.* Washington, D.C.: National Academy Press.

Esty, S. A. (1987). "Fiber to the home" activity in the United States of America. In *IEEE/IEICE Global Telecommunications Conference 1987 Conference Record* (Vol. 3, pp. 1995-1999). Washington, DC: IEEE.

Federal Communications Commission. (2004). *High-speed services for Internet access: Status as of December 31, 2003.* Retrieved June 18, 2004, from *http://www.fcc.gov/Bureaus/Common_Carrier/Reports/FCC-State_Link/IAD/hspd0604.pdf.* Washington, D.C.: IEEE

IEEE. (2002). *IEEE standard 802.16-2001*. New York: IEEE.

IEEE. (2003). *IEEE standard 802.16a-2003*. New York: IEEE.

IEEE 802.3ah Ethernet in the First Mile Task Force. (2004). *Draft of IEEE P802.3ah*. New York: IEEE.

The International Engineering Consortium (IEC). (n.d.). *Very-high-data-rate digital subscriber line (VDSL)*. Chicago: The International Engineering Consortium. Retrieved July 7, 2004, from *http://www.iec.org/online/tutorials/vdsl/*

Ismail, S., & Wu, I. (2003, October). *Broadband Internet access in OECD countries: A comparative analysis.* Washington, D.C.: FCC Retrieved July 7, 2004, from *http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-239660A2.pdf*

Rowbotham, T. R. (1989). Plans for a British trial of fibre to the home. *British Telecommunications Engineering, 8*(2), 78-82.

Shumate, P. W., Jr. (1989, February). Optical fibers reach into homes. *IEEE Spectrum*, *26*(2), 43-47. New York: IEEE.

Sirbu, M., & Reed, D. (1988). An optimal investment strategy model for fiber to the home. In *Proceedings: International Symposium on Subscriber Loops and Services, ISSLS 88* (pp. 149-155). New York: IEEE.

Telecommunication Standardization Sector of ITU (ITU-T). (1998). *ITU-T recommendation G.983.1: Broadband optical access systems based on passive optical networks (PONs).* Geneva, Switzerland: International Telecommunication Union.

Telecommunication Standardization Sector of ITU (ITU-T). (2001). *ITU-T recommendation G.983.3: A broadband optical access system with increased service capability by wavelength allocation.* Geneva, Switzerland: International Telecommunication Union.

Telecommunication Standardization Sector of ITU (ITU-T). (2003a). *ITU-T recommendation G.984.1: Gigabit-capable passive optical networks (GPON): General characteristics.* Geneva, Switzerland: International Telecommunication Union.

Telecommunication Standardization Sector of ITU (ITU-T). (2003b). *ITU-T recommendation G.984.2: Gigabit-capable passive optical networks (GPON): Physical media dependent (PMD) layer specification.* Geneva, Switzerland: International Telecommunication Union.

Telecommunication Standardization Sector of ITU (ITU-T). (2003c). *ITU-T recommendation G.985: 100 Mbit/s point-to-point Ethernet based optical access system.* Geneva, Switzerland: International Telecommunication Union.

Telecommunication Standardization Sector of ITU (ITU-T). (2004). *ITU-T recommendation G.984.3: Gigabit-capable passive optical networks (GPON): Transmission convergence layer specification.* Geneva, Switzerland: International Telecommunication Union.

Veyres, C., & Mauro, J. J. (1988). Fiber to the home: Biarritz (1984)…twelve cities (1988). In *IEEE International Conference on Communications, 1988: Digital Technology—Spanning the Universe* (Vol. 2, pp. 874-888). New York: IEEE.

KEY TERMS

APON or Broadband PON (APON/BPON): APON is defined by the ITU-T G.983 series of recommendations. It features a passive optical network for fiber-to-the-home service that uses ATM as its transmission protocol. BPON is an alternate name for this technology.

Broadband: The U.S. Federal Communications Commission defines broadband to be any high-speed digital technology that provides integrated access to high-speed data, video-on-demand, and interactive delivery services with a data rate of at least 200 kbps in one direction.

EFM Fiber: EFM Fiber is defined by IEEE 802.3ah. It features a point-to-point fiber-to-the-home network, typically deployed as an active star, that uses active electronics and Ethernet as its transmission protocol.

Ethernet PON (EPON): EPON is defined by IEEE 802.3ah. It features a passive optical network for fiber-to-the-home service that uses Ethernet as its transmission protocol.

Fiber-to-the-Home (FTTH): The use of fiber-optic cable for the provisioning of narrowband and broadband services to the residential customer rather than traditional copper wiring.

Gigabit PON (GPON): GPON is defined by the ITU-T G.984 series of recommendations. It features a passive optical network for fiber-to-the-home ser-

vice that is capable of providing at least 1 Gbps service in the downstream direction.

Narrowband: A transmission path that is capable of 64 kbps transmission and voice-grade service.

Optical Line Terminal (OLT): A fiber-to-the-home terminating device at the provider's central office or point of presence connected to one or more PONs that provides connection to the provider's network.

Optical Network Terminal (ONT): A fiber-to-the-home terminating device at the customer premise.

Passive Optical Network (PON): An optical transmission path from the provider to the customer that contains only unpowered optical components, such as optical splitters and splices.

From Communities to Mobile Communities of Values

Patricia McManus
Edith Cowan University, Australia

Craig Standing
Edith Cowan University, Australia

INTRODUCTION

The discussion around the impact of information communication technologies in human social interaction has been the centre of many studies and discussions. From 1960 until 1990, researchers, academics, business writers, and futurist novelists have tried to anticipate the impact of these technologies in society, in particular, in cities and urban centres (Graham, 2004). The views during these three decades, although different in many aspects, share in common a deterministic view of the impact of ICT on cities and urban centres. They all see ICT influence as a dooming factor to the existence of cities. These authors have often seen ICT as a leading factor in the disappearance of urban centres and/or cities (Graham; Marvin, 1997; Negroponte, 1995). According to Graham, these views tend to portray ICT impact without taking into consideration the fact that old technologies are not always replaced by newer ones; they can also superimpose and combine into to something else. These views also have generally assumed that the impact of ICT would be the same in all places and have not accounted for geographic differences that could affect the use of information communication technologies.

This article assesses the significance of the theory of consumption value as an explanatory framework for mobile commerce (m-commerce) adoption and use. It discusses whether perceived values can define the characteristics of any discrete "community of use" (group) of m-commerce users. It discusses the significance of online communities and their relation with mobile commerce. We first discuss the impact of ICT in cities. Second, we present the theory of consumption values as a framework to understand mobile commerce use. Then we assess the relevance of communities' values as an explanatory theory to mobile commerce adoption. Finally, we explore the possibility that consumption values could be mobile-community-binding instruments.

There are a few weaknesses in these deterministic views of the impact of ICT on the development or dooming of cities. Most of them assume that technology impacts exactly the same way everywhere; that is, there is an assumption that a city is the same anywhere on the globe (Graham, 2004). This perspective, also, does not take into account the growth of physical mobility in urban centres (Graham) and the fact that technology does not promote only isolationism (Horan, 2004). Statistics show, for example, that there was a continuous rise in global motor vehicle ownership, from 350 million in 1980 to 500 million in 2001, and a forecast of 1 billion by 2030 (Bell & Gemmel, 2001). Moreover, "in 2001 more mobile phones were shipped than automobiles and PCs" (Clarke, 2001, p. 134). In 200l, out of the 200 million wireless devices sold in the U.S., 13.1 million were personal digital assistants (PDAs) and the other 187 million were mobile phones (Strauss, El-Ansary, & Frost, 2003). It is important, though, not to presume that some level of face-to-face contact is not going to be replaced by electronic technology. Refer, for example, to what is happening with many network-based services like online banking, EDI (electronic data interchange), or the DoCoMo phenomenon in Japan (Graham; Krishnamurthy, 2001). It becomes reasonable to assume that it is very unlikely that ICTs will bring death to the cities. On the contrary, they are deeply entrenched in urbanisation and social economic trends (Graham).

RELEVANCE OF COMMUNITIES

Many works in cultural geography, sociology, and anthropology refer to the mediating role of technologies in structuring the relationship between individuals and their social environment or community (Green, 2002).

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Community can be defined as "the formation of relatively stable long-term online group associations" (Barkardjiva & Feenberg, 2002, p. 183). Traditionally, the concept of community is associated with many circumstances or factors; however, a common physical location was for many years considered to be a key factor to determine their existence (Graham, 2004). With the development and popularization of ICTs, in particular, the Internet and mobile phones, it is possible to say that the key factor to determine the existence of a community is accessibility (Webber, 2004)

In the social sciences, the concept of community has generated so much discussion that it has already reached a theoretical sophistication (Komito, 1998). However, this theoretical sophistication has not been transferred to the concept of ICT-mediated communities (Komito). The broad interpretation of the community concept in the network environment has many different meanings, ranging from definitions like "norm or values shared by individuals," "a loose collection of like-minded individuals," or "a multifaceted social relation that develops when people live in the same locality and interact, involuntarily, with each other over time" (Komito, p. 97). We consider virtual communities to refer to different types of communities facilitated by information communication technology.

Authors Armstrong and Hagel (1999) were two of the pioneers in using the term virtual community. By virtual community they describe a group of technology enthusiasts in San Francisco. These high-tech enthusiasts created a space in the early days of the Internet prior the World Wide Web. This was and still is a site where people can get together to discuss and exchange cultural information, and today it has migrated to the Web. "The well has been a literate watering hole for thinkers from all walks of life, be they artists, journalists, programmers, educators or activists" (The Well, 2003). Haylock and Muscarella (1999) on the other hand, use the term virtual community when referring specifically to the World-Wide-Web-based communities, but kept their definition of community quite broad. To them a virtual community is a "group of individuals who belong to particular demographic, profession or share a particular personal interest" (p. 73).

In his 1998 article, Komito discusses extensively the community concept and develops a taxonomy for virtual and electronic communities. He identifies three basic kinds of communities: the moral community (the character of the social relationship is paramount), normative or cognitive community (existence of preset rules of behaviour), and proximate community (interaction happens not because of roles or stereotypes, but because of individuals). A moral community refers to people who share a common ethical system, and it is this shared ethical system that identifies their members. According to Komito, this kind of community is difficult to identify in a computer-mediated communication environment, with the moral purpose of the community being difficult to identify. The normative community is probably the most common type of community associated with ICT. This kind of community is not bound physically or geographically, but is bound by common meaning and culture, such as members being medical doctors, Jews, or jazz aficionados. The individual participants in these communities may never interact with all the other members of this particular community. Authors such as Komito believe that the concepts of community of interest and community of practice borrowed their framework from cognitive communities. Proximate communities have a social emphasis. In this model of community, the interaction between members happens not only in terms of roles or stereotypes, but at the individual level; it is in this kind of community where relationships are developed and conflicts managed (Komito). Although he presented a typology for ICT-mediated communities, Komito concludes that the most useful way of looking at ICT-mediated communities would be to treat the community as a background and concentrate on how individuals and groups deal and adapt to continuously changing environments in terms of social interaction rules. With this in mind, we suggest that a group of individuals who share the same consumption values in relation to mobile services could be members of the same community. The concept of consumption values comes from Sheth, Newman, and Gross' (1991a, 1991b) theory, described next.

THEORY OF CONSUMPTION VALUES: AN ALTERNATIVE FRAMEWORK TO UNDERSTAND MOBILE COMMERCE USE

In reviewing the literature on the adoption and use of technologies, some dominant theoretical frameworks were identified as adaptations or extensions to Rogers' (1962, 2003) diffusion-of-innovation theory

or Ajzen's (1991) theory of planned behaviour (TPB). The technology-adoption model (TAM; Davis, Zaner, Farnham, Marcjan, & McCarthy, 1989) is derived from Ajzen and Fishbein's (1980) theory of reasoned action (TRA; which TPB is based upon). Most recently, Venkatesh, Morris, Davis, and Davis (2003) conceptualized the unified theory of acceptance and use of technology (UTAUT). This model is quite comprehensive as it combines TRA, TAM, TPB, the IDT (Innovation Diffusion Theory) model of MPCU (Model of PC Utilization) (personal computer) utilization, the motivational model, and social cognitive theory. However, as the model integrates several theories that focus on user and consumer intention to behave, this model does not concentrate on actual behaviour. For this reason we suggest the utilization of Sheth et al.'s (1991a) theory of consumption values. Although this model has not been directly applied to technology adoption, its unique perspective on consumption values can provide valuable insights to better understand m-commerce-adoption drivers.

Sheth et al. (1991a, 1991b) conceptualized a model to help comprehend how consumers make decisions in the marketplace. They based their model on the principle that the choices consumers make are based on their perceived values in relation to what the authors called "market choice," and that the perceived values contribute distinctively to specific choices. Because their model examines what the product values are that attract consumers, it can be viewed as a way to understand the attitude toward the product, making this a proactive way to understand m-commerce adoption.

Sheth et al. (1991a) classify five categories of perceived value. Functional values are associated with the utility level of the product (or service) compared to its alternatives. Social value is described as the willingness to please, and social acceptance. Emotional values are those choices made based upon feelings and aesthetics. A common example would be the choice of sports products. Epistemic values can be used to describe the early adopters in the sense that it relates to novelty or knowledge-searching behaviour. Words such as *cool* and *hot* are often associated with this value. Finally, the conditional value refers to a set of circumstances depending on the situation (e.g., Christmas, a wedding, etc.). Socioeconomical and physical aspects are included in this value. These five values were conceptualized based on a diversity of disciplines including social psychology, clinical psychology, sociology, economics, and experimental psychology (Sheth et al., 1991a).

This theory has not been used to directly explain adoption; however, its unique conceptualization of product values provides a multidisciplinary approach that would contribute toward the understanding of the actual consumer behaviour in a market choice situation. The limitation of this theory to understanding adoption is that it cannot be used to understand organisational adoption as it does not address influential factors that affect purchase couples or group adoption. Another limitation is that this model cannot be used to understand adoption in cases where the buyer is not the user. Nevertheless, Sheth et al.'s model (1991a) "provides the best foundation for extending value construct as it was validated through an intensive investigation in a variety of fields in which value has been discussed" (Sweeney & Soutar, 2001, p. 205).

The application of Sheth et al.'s model (2001a) would help to provide an understanding of intrinsic influential factors, that is, values about electronic channels such as mobile services (Amit & Zott, 2001; Anckar, 2002; Eastilick & Lotz, 1999; Han & Han, 2001; Venkatesh & Brown, 2001). The theory of consumption values can identify the main value-adding elements in m-commerce or the primary drivers for adopting m-commerce.

Sheth et al. (1991a, 1991b) claim that the main limitation of the theory of consumption value is the fact that it cannot be used to predict the behaviour of two or more individuals. However, this may not be true if the individuals form a group because they share the same perceived values.

COMMUNITIES OF VALUE

The community concept has been used in a number of areas in information systems research. The emergence of networked technologies and the popularization of the Internet have brought a new approach to the study of communities (Bakardjiva & Feenberg, 2000; Haylock and Muscarella, 1999; Komito, 1998). Authors have used the terms online community and virtual community interchangeably. However, one can say that the term virtual community is far broader and may include any technology-mediated communi-

cation, whilst online community would be more applicable to the Internet or the World-Wide-Web portion of the Internet. Also, communities of practice have been in the centre of academic journals' and practitioners' publications' attention; however, this community is not dependent on technology. In fact, they have been around for centuries. They can be defined "as groups of individuals informally bound together by shared expertise and passion for a shared enterprise" (Wenger & Snyder, 2000, p. 139). When studying virtual communities, researchers seek to understand and classify the role that network technology plays in structuring relationships, societies, and their subsets (Armstrong & Hagel, 1999; Bakardjiva & Feenberg; Haylock & Muscarella, 1999). The interest on communities of practice has been driven by researchers who have identified these informal, self-organised nodes. These groups have been identified as beneficial to organisations, and their strength lies in their ability to self-perpetuate and generate knowledge (Wenger & Snyder).

In information systems, studies of communities have helped to better understand systems adoption and usability. In marketing, communities are now an alternative way to segment consumers (Table 1). Mobile technologies have had a profound impact on people's everyday lives to the point of reshaping time and space (Green, 2002). Green explores the impact of mobile technologies in time and space. Underpinning her arguments are concepts such as proximity, mobile work, flexible schedules, and so forth, which depict this new understanding of temporality. In today's life, social relationships have become fragmented, and mobile technologies represent a way to bring continuity back (Green). This new mobile lifestyle is quite prevalent in teenagers. Spero's (2003) white paper points out that the old demographic segmentation of teenagers (ages seven to 10 as tweens, 11 to 13 as young teens, 14 to 16 as teenagers, and 16 and older as young adults) is no longer effective, and a more efficient alternative is segmentation based on mobile lifestyle. These lifestyle traits encompass things like interest, behaviour, upbringing, and eating habits. We propose that identifying communities of mobile service value through the underlying reasons why users perceive those values, from Sheth et al.'s (1991a, 1991b) theory, provides a theoretical framework for understanding mobile service adoption.

CONCLUSION

There are great expectations in relation to the adoption of m-commerce. This article has discussed the utilization of the theory of consumption value (Sheth et al., 1991a, 1991b) as an alternative framework to understand m-commerce adoption and use. The value theory provides deeper explanatory ability as it examines the underlying rationale in the decision-making process. This can more easily be used for predictive purposes. For example, a main driver for teenagers using mobile phones is the relatively low cost of text messaging; however, the motivator for use is the intrinsic social aspect of the service, which caters and builds upon an existing community of use.

Product and service developers need to examine these deeper factors to come to a sophisticated understanding of adoption-related decisions. Previous theoretical explanations for technology adoption are low in terms of predictive capabilities. This article suggests that the consumer perceived-values approach has significant potential not only in explaining adoption decisions on an individual level, but also across communities of use or practice. These communities exist in the business world as well as society in general.

The concept of community of use represents a more effective way to identify different groups or segments as demographics are no longer reliable. People within the same age group do not necessarily have the same lifestyle and perceive the same values in a service.

Table 2. Examples of communities of use

Community of Use	Lifestyle (Common Traits)	Dominant Perceived value	Issues within the Values	Type of Service
Nomadic Professional	Virtual Office	Functional	Convenience	Micropayment (Parking)
Urban Teens Social Group	Connected Net Generation Sociable	Social	Short Messages	SMS
Postmodern Family	Discontinuous	Functional	Convenience	Voice, SMS

The value perceived in a service or product could be what binds groups of individuals in communities, generating what one would call communities of values. The reasons why individuals perceive some values in mobile services can explain group behaviour.

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and the Human Decision Process, 50*, 179-211.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Englewood Cliffs, NJ: Prentice-Hall.

Amit, R., & Zott, C. (2001). Value creation in e-business. *Strategic Management Journal, 22*, 493-520.

Anckar, B. (2002). Adoption drivers and intents in the mobile electronic marketplace: Survey findings. *Journal of System and Information Technology, 6*(2), 1-17.

Armstrong, A., & Hagel, J., III. (1999). The real value of online communities. In D. Tapscott (Ed.), *Creating value in the network economy* (pp. 173-185). Boston: Harvard Business School Publishing.

Bakardjiva, M., & Feenberg, A. (2002). Community technology and democratic rationalization. *The Information Society, 18*(3), 181-192.

Bell, G., & Gemmell, J. (2002). A call for the home media network. *Communications of the ACM, 45*(7), 71-75.

Brown, K. M. (1999). *Theory of reasoned action/ theory of planned behaviour*. University of South Florida. Retrieved June 21, 2003, from *http://hsc.usf.edu/ ~kmbrown/TRA_TPB.htm*

Clarke, I., III. (2001). Emerging value propositions for m-commerce. *Journal of Business Strategies, 18*(2), 133-148.

Davis, J., Zaner, M., Farnham, S., Marcjan, C., & McCarthy, B.P. (2003, January 7-10). *Wireless brainstorming: Overcoming status effects in small group decisions.* Paper presented at the 36th Hawaii International Conference on Systems Sciences, Big Island, Hawaii.

Eastlick, M. A, & Lotz, S. (1999). Profiling potential adopters of interactive teleshopping. *International Journal of Retail and Distribution Management, 27*(6), 209-228.

Fano, A., & Gershman, A. (2002). The future of business services. *Communications of the ACM, 45*(12), 83-87.

Graham, S. (2004). Introduction: From dreams of transcendence to the remediation of urban life. In S. Graham (Ed.), *The cybercities reader* (pp. 1-33). London: Routledge Taylor & Francis Group.

Green, N. (2002). On the move: Technology, mobility, and the mediation of social time and space. *The Information Society, 18*(3), 281-292.

Han, J., & Han, D. (2001). A framework for analysing customer value of Internet business. *Journal of Information Technology Theory and Application (JITTA), 3*(5), 25-38.

Han, S., Harkke, V., Landor, P., & Mio, R. R. d. (2002). A foresight framework for understanding the future of mobile commerce. *Journal of Systems & Information Technology, 6*(2), 19-39.

Haylock, C., & Muscarella, L. (1999). Virtual communities. In C. Haylock & L. Muscarella (Eds.), *Net success* (chap. 4, p. 320). Holbrook, MA: Adams Media Corporation.

Ho, S. Y., & Kwok, S. H. (2003). The attraction of personalized service for users in mobile commerce: An empirical study. *ACM SIGecom Exchanges, 3*(4), 10-18.

Horan, T. (2004). Recombinations for community meaning. In S. Graham (Ed.), *The cybercities reader*. London: Routledge, Taylor & Francis Group.

Jackson, P. B., & Finney, M. (2002). Negative life events and psychological distress among young adults. *Social Psychology Quarterly, 65*(2), 186-201.

Klein, H. K., & Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly, 23*(1), 67-94.

Komito, L. (1998). The Net as a foraging society: Flexible communities. *The Information Society, 14*(2), 97-106.

Krishnamurthy, S. (2001). *NTT DoCoMo's I-Mode phone: A case study*. Retrieved March 17, 2003, from *http://www.swcollege.com/marketing/krishnamur thy/ first_edition/ case_updates/docomo_final.pdf*

Levy, M. (2000). Wireless applications become more common. *Commerce Net.* Retrieved July 5, 2003, from *http://www.commerce.net/research/ebusiness-strategies/2000/00_13_n.html*

Marvin, S. (1997). Environmental flows: Telecommunications and dematerialisation of cities. *Futures, 29*(1).

Negroponte, N. (1995). *Being digital.* London: Hodder & Stoughton.

Rogers, E.M. (2003). Diffusion of innovations (5th ed.). New York: Free Press, A division of Simon & Schuster, Inc. 1230 Avenue. (1962 - 1st ed.).

Ropers, S. (2001, February). New business models for the mobile revolution. *EAI*, 53-57. Available at *http://www.bijonline.com/PDF/Mobile%20Revolution% 20-%20Ropers.pdf*

Sheth, J. N., Newman, B. I., & Gross, B. L. (1991a). *Consumption values and market choice: Theory and applications.* Cincinnati, OH: South-Western Publishing Co.

Sheth, J. N., Newman, B. I., & Gross, B. L. (1991b). Why we buy what we buy: A theory of consumption values. *Journal of Business Research, 22*, 150-170.

Spero, I. (2003). *Agents of change. Teenagers: Mobile lifestyle trends*. Retrieved November 28, 2003, from *http://www.spero.co.uk/agentsofchange*

Strauss, J., El-Ansary, A., & Frost, R. (2003). *E-marketing* (3rd ed.). Upper Saddle River, NJ: Pearson Education Inc.

Sweeney, J. C., & Soutar, G. N. (2001). Consumer perceived value: The development of a multiple item scale. *Journal of Retailing, 77*(2), 203-220.

Sweeney, J. C., Soutar, G. N., & Johnson, L. W. (1999). The role of perceived risk in the quality-value relationship: A study in a retail environment. *Journal of Retailing, 77*(1), 75-105.

Tierney, W. G. (2000). Undaunted courage: Life history and the postmodern challenge. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 537-554). Thousand Oaks, CA: Sage.

Venkatesh, V., & Brown, S. A. (2001). A longitudinal investigation of personal computers in homes: Adoption determinants and emerging challenges. *MIS Quarterly, 25*(1), 71-102.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478.

Webber, M. (2004). The urban place and the non-place urban realm. In S. Graham (Ed.), *The cybercommunities reader* (pp. 50-56). London: Routledge.

The Well. (2003). Retrieved November 25, 2003, from *http://www.well.com/aboutwell.html*

Wenger, E. C., & Snyder, W. M. (2000, January-February). Communities of practice: The organizational frontier. *Harvard Business Review*, (January-February), 139-145.

KEY TERMS

DoCoMo: Japanese mobile telecommunication company that is a part of NTT. It is the creator of I-Mode.

EDI (Electronic Data Interchange): A set of computer interchange standards developed in the '60s for business documents such as invoices, bills, and purchase orders. It has evolved to use the Internet.

TAM (Technology-Acceptance Model): Described as an adaptation of TRA customised to technology acceptance. The intention to adopt is affected by two beliefs: perceived usefulness and the perceived ease of use of the new technology.

TPB (Theory of Planned Behaviour): TPB is an extension of TRA. It adds a third dimension—the perceived-behaviour control component—that looks at uncontrolled external circumstances.

TRA (Theory of Reasoned Action): TRA states that the intention to adopt is affected directly by attitudinal components (beliefs about the outcome of the behaviour and beliefs of the consequences of the behaviour) and the subjective norm component (level of importance or desire to please significant others and/or society).

UTAUT (Unified Theory of Acceptance and Use of Technology): This model is quite comprehensive as it combines TRA, TAM, TPB, the DOI model of PC utilization, the motivational model, and social cognitive theory.

The Future of M-Interaction

Joanna Lumsden
National Research Council of Canada IIT e-Business, Canada

INTRODUCTION

Many experts predicted that this, the first decade of the 21st century, will be the decade of mobile computing; although in recent years mobile technology has been one of the major growth areas in computing, the hype has thus far exceeded the reality (Urbaczewski, Valacich, & Jessup, 2003). Why is this? A recent international study of users of handheld devices suggests that there is a predominant perception that quality of service is low and that mobile applications are difficult to use; additionally, although users recognise the potential of emerging mobile technology, the study highlighted a general feeling that the technology is currently dominating rather than supporting users (Jarvenpaa, Lang, Takeda, & Tuunainen, 2003). Users are generally forgiving of physical limitations of mobile devices imposed by technological constraints; they are not, however, so forgiving of the interface to these devices (Sarker & Wells, 2003). Users can excuse restrictions on their use of mobile technology on the basis of level of technological advancement, but find it hard to accept impractical, illogical, or inconvenient interaction design.

Mobile devices are becoming increasingly diverse and are continuing to shrink in size and weight. Although this increases the portability of such devices, their usability tends to suffer. Screen sizes are becoming smaller making them hard to read. If interaction design for mobile technologies does not receive sufficient research attention, the levels of frustration—noted to be high for mobile technology and fuelled almost entirely by lack of usability (Venkatesh, Ramesh, & Massey, 2003)—currently experienced by m-commerce users will only worsen. Widespread acceptance of mobile devices amongst individual consumers is essential for the promise and commercial benefit of mobility and m-commerce to be realised. This level of acceptance will not be achieved if users' interaction experience with mobile technology is negative. We have to design the right types of m-interaction if we are to make m-commerce a desirable facility in the future; an important prerequisite for this is ensuring that users' experience meets both their sensory and functional needs (Venkatesh et al., 2003).

Given the resource disparity between mobile and desktop technologies, successful e-commerce interface design does not necessarily equate to successful m-commerce design. It is therefore imperative that the specific needs of m-commerce are addressed in order to heighten the potential for acceptance of m-commerce as a domain in its own right. This chapter begins by exploring the complexities of designing interaction for mobile technology, highlighting the effect of context on the use of such technology. It then goes on to discuss how interaction design for mobile devices might evolve, introducing alternative interaction modalities that are likely to affect that future evolution. By highlighting some of the possibilities for novel interaction with mobile technology it is hoped that future designers will be encouraged to "think out of the box" in terms of their designs and, by doing so, achieve greater levels of acceptance of m-commerce.

THE COMPLEXITY OF DESIGNING INTERACTION FOR MOBILITY

Despite the obvious disparity between desktop systems and mobile devices in terms of "traditional" input and output capabilities, the user interface designs of most mobile devices are based heavily on the tried-and-tested desktop design paradigm. Desktop user interface design originates from the fact that users are stationary—that is, seated at a desk—and can devote all or most of their attentional resources to the application with which they are interacting. Hence, the interfaces to desktop-based applications are typically very graphical (often very detailed) and use the standard keyboard and mouse to facilitate interaction. This has proven to be a very

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

successful paradigm which has been enhanced by the availability of ever more sophisticated and increasingly larger displays.

Contrast this with mobile devices—for example, cell phones, personal digital assistants (PDAs), and wearable computers. Users of these devices are typically in motion when using their device. This means that they cannot devote all of their attentional resources—especially visual resources—to the application with which they are interacting; such resources must remain with their primary task, often for safety reasons (Brewster, 2002). Additionally, mobile devices have limited screen real estate and standard input and output capabilities are generally restricted. This makes designing mobile interaction (m-interaction) difficult and ineffective if we insist on adhering to the tried-and-tested desktop paradigm. Poor m-interaction design has thus far led to disenchantment with m-commerce applications: m-interaction that is found to be difficult results in wasted time, errors, and frustration that ultimately end in abandonment.

Unlike the design of interaction techniques for desktop applications, the design of m-interaction techniques has to address complex contextual concerns. Sarker and Wells (2003) identify three different modes of mobility—travelling, wandering, and visiting—which they suggest each motivate use patterns differently. Changing modality of mobility is actually more complex than simply the reason for being mobile: with mobility comes changes in several different contexts of use.

Most obviously, the physical context in which the user and technology is operating constantly changes as the user moves. This includes, for example, changes in ambient temperatures, lighting levels, noise levels, and privacy implications. Connected to changing physical context is the need to ensure that a user is able to safely navigate through his/her physical environment while interacting with the mobile technology. This may necessitate m-interaction techniques that are eyes-free and even hands-free. This is not a simple undertaking given that such techniques must be sufficiently robust to accommodate the imprecision inherent in performing a task while walking, for example.

Users' m-interaction requirements also differ based on task context. Mobile users inherently exhibit multitasking behaviour which places two fundamental demands on m-interaction design: firstly, interaction techniques employed for one task must be sympathetic to the requirements of other tasks with which the user is actively involved— for instance, if an application is designed to be used in a motor vehicle, for obvious safety reasons, the m-interaction techniques used cannot divert attention from the user's primary task of driving; secondly, the m-interaction technique that is appropriate for one task may be inappropriate for another task—so, unlike the desktop paradigm, we cannot adopt a one-technique-fits-all approach to m-interaction.

Finally, we must take the social context of use into account when designing m-interaction techniques; if we are to expect users to wear interaction components or use physical body motion to interact with mobile devices, at the very least we have to account for social acceptance of behaviour. In actual fact, the social considerations relating to use of mobile technology extend beyond behavioural issues; however, given the complexity of this aspect of technology adoption (it is a research area in its own right) it is beyond the immediate scope of this discussion. That said, it is important to note that technology that is not, at its inception, considered socially acceptable, can gain acceptability with usage thresholds and technological evolution—consider, for example, acceptance of cell phones.

EVOLVING INTERACTION DESIGN FOR MOBILITY

The great advantage the telephone possesses over every other form of electrical apparatus consists in the fact that it requires no skill to operate the instrument. Alexander Graham Bell, 1878

The above observation from Alexander Graham Bell, the founder of telecommunications, epitomises what we must hold as our primary goal when designing future m-interaction; that is, since the nature of mobile devices is such that we cannot assume users are skilled, m-interaction should seem natural and intuitive and should fit so well with mobile contexts of use that users feel no skill is required to use the associated mobile device. Part of achieving this is acquiring a better understanding of the way in which mobility affects the use of mobile devices and thereafter designing m-interaction to accommodate these

influences. Additionally, we need to better understand user behaviour and social conventions in order to align m-interaction with these key influences over mobile device use. Foremost, we need to design m-interaction such that a mix of different interaction styles are used to overcome device limitations (for example, screen size restrictions). Ultimately, the key to success in a mobile context will be the ability to present, and allow users to interact with, content in a customized and customizable fashion.

It is hard to design purely visual interfaces that accommodate users' limited attention; that said, much of the interface research on mobile devices tends to focus on visual displays, often presented through head-mounted graphical displays (Barfield & Caudell, 2001) which can be obtrusive, are hard to use in bright daylight, and occupy the user's visual resource (Geelhoed, Falahee, & Latham, 2000). By converting some or all of the content and interaction requirements from the typical visual to audio, the output space for mobile devices can be dramatically enhanced and enlarged. We have the option of both speech and non-speech audio to help us achieve this.

Speech-Based Audio

Using voice technologies, users issue commands to a system simply by speaking, and output is returned using either synthesised or pre-recorded speech (Beasly, Farley, O'Reilly, & Squire, 2002; Lai & Yankelovich, 2000). Voice-based systems can use constrained (Beasly et al., 2002) or unconstrained (Lai & Yankelovich, 2000) vocabularies with accordingly different levels of sophistication balanced against accuracy. This type of m-interaction can seem very natural; it can permit eyes-free and even hands-free interaction with m-commerce applications. However, perhaps more so than any of the other possible m-interaction techniques, speech-based interaction faces a number of environmental hurdles: for instance, ambient noise levels can render speech-based interaction wholly impractical and for obvious reasons, privacy is a major concern. When used for both input and output, speech monopolises our auditory resource—we can listen to non-speech audio while issuing speech-based commands, but it is hard to listen to and interpret speech-based output while issuing speech-based input. That said, given appropriate contextual settings, speech-based interaction—especially when combined with other interaction techniques—is a viable building block for m-interaction of the future.

Non-Speech Audio

Non-speech audio has proven very effective at improving interaction on mobile devices by allowing users to maintain their visual focus on navigating through their physical environment while presenting information to them via their audio channel (Brewster, 2002; Brewster, Lumsden, Bell, Hall, & Tasker, 2003; Holland & Morse, 2001; Pirhonen, Brewster, & Holguin, 2002; Sawhney & Schmandt, 2000).

Non-speech audio, which has the advantage that it is language independent and is typically fast, generally falls into two categories: "earcons", which are musical tones combined to convey meaning relative to application objects or activities, and "auditory icons", which are everyday sounds used to represent application objects or activities. Non-speech audio can be multidimensional both in terms of the data it conveys and the spatial location in which it is presented. Most humans are very good at streaming audio cues, so it is possible to play non-speech audio cues with spatial positioning around the user's head in 3D space and for the user to be able to identify the direction of the sound source and take appropriate action (for example, selecting an audio-representation of a menu item). Non-speech audio clearly supports eyes-free interaction, leaving the speech channel free for other use. However, non-speech audio it is principally an output or feedback mechanism; to be used effectively within the interface to mobile devices, it needs to be coupled with an input mechanism. As intimated previously, speech-based input is a potential candidate for use with non-speech audio output; so too, however, is gestural input.

Audio-Enhanced Gestural Interaction

Gestures are naturally very expressive; we use body gestures without thinking in everyday communication. Gestures can be multidimensional: for example, we can have 2D hand-drawn gestures (Brewster et al., 2003; Pirhonen et al., 2002), 3D hand-generated gestures (Cohen & Ludwig, 1991), or even 3D head-generated gestures (Brewster et al., 2003). Harrison, Fishkin, Gujar, Mochon, and Want (1998) showed

that simple, natural gestures can be used for input in a range of different situations on mobile devices. Head-based gestures are already used successfully in software applications for disabled users; as yet, however, their potential has not been fully realised nor fully exploited in other applications. There has, until recently, been little use of audio-enhanced physical hand and body gestures for input on the move; such gestures are advantageous because users do not need to look at a display to interact with it (as they must do, for example, when clicking a button on a screen in a visual display). The combined use of audio and gestural techniques present the most significant potential for viable future m-interaction. Importantly, gestural and audio-based interaction can be eyes-free and, assuming non hand-based gestures, can be used to support hands-free interaction where necessary.

A seminal piece of research that combines audio output and gestural input is Cohen and Ludwig's *Audio Windows* (Cohen & Ludwig, 1991). In this system, users wear a headphone-based 3D audio display in which application items are mapped to different areas in the space around them; wearing a data glove, users point at the audio represented items to select them. This technique is powerful in that it allows a rich, complex environment to be created without the need for a visual display—important when considering m-interaction design. Savidis, Stephanidis, Korte, Crispien, and Fellbaum also developed a non-visual 3D audio environment to allow blind users to interact with standard GUIs (Savidis et al., 1996); menu items are mapped to specific places around the user's head and, while seated, the user can point to any of the audio menu items to make a selection. Although neither of these examples was designed to be used when mobile, they have many potential advantages for m-interaction.

Schmandt and colleagues at MIT have done work on 3D audio in a range of different applications. One, *Nomadic Radio*, uses 3D audio on a mobile device (Sawhney & Schmandt, 2000). Using non-speech and speech audio to deliver information and messages to users on the move, *Nomadic Radio* is a wearable audio personal messaging system; users wear a microphone and shoulder-mounted loudspeakers that provide a planar 3D audio environment. The 3D audio presentation has the advantage that it allows users to listen to multiple sound streams simultaneously while still being able to distinguish and separate each one (the "Cocktail Party" effect). The spatial positioning of the sounds around the head also conveys information about the time of occurrence of each message.

Pirhonen et al. (2002) examined the effect of combining non-speech audio feedback and gestures in an interface to an MP3 player on a Compaq iPAQ. They designed a small set of metaphorical gestures, corresponding to the control functions of the player, which users can perform, while walking, simply by dragging their finger across the touch screen of the iPAQ; users receive end-of-gesture audio feedback to confirm their actions. Pirhonen et al. (2002) showed that the audio-gestural interface to the MP3 player is significantly better than the standard, graphically-based media player on the iPAQ.

Brewster et al. (2003) extended the work of Pirhonen et al. (2002) to look at the effect of providing non-speech audio feedback during the course of gesture generation as opposed to simply providing end-of-gesture feedback. They performed a series of experiments during which participants entered, while walking, alphanumeric and geometrical gestures using a gesture recogniser both with and without dynamic audio feedback. They demonstrated that by providing non-speech audio feedback during gesture generation, it is possible to improve the accuracy—and *awareness* of accuracy—of gestural input on mobile devices when used while walking. Furthermore, during their experiments they tested two different soundscape designs for the audio feedback and found that the simpler the audio feedback design the better to reduce cognitive demands placed upon users.

Fiedlander, Schlueter and Mantei (1998) developed non-visual "Bullseye" menus where menu items ring the user's cursor in a set of concentric circles divided into quadrants. Non-speech audio cues—a simple beep played without spatialisation—indicate when the user moves across a menu item. A static evaluation of Bullseye menus showed them to be an effective non-visual interaction technique; users are able to select items using just the sounds. Taking this a stage further, Brewster et al. (2003) developed a 3D auditory radial pie menu from which users select menu items using head nods. Menu items are displayed in 3D space around the user's head at the level of the user's ears and the user selects an item by nodding in the direction of the item. Brewster et al. (2003) tested three different soundscapes for the

presentation of the menu items, each differing in terms of the spatial positioning of the menu items relative to the user's head. They confirmed that head gestures are a viable means of menu selection and that the soundscape that was most effective placed the user in the middle of the menu, with items presented at the four cardinal points around the user's head.

CONCLUSION

The future of m-interaction looks exciting and bright if we embrace the possibilities open to us and adopt a paradigm shift in terms of our approach to user interface design for mobile technology. This discussion has highlighted some of those possibilities, stressing the potential for combined use of audio and gestural interaction as it has been shown to be an effective combination in terms of its ability to significantly improve the usability of mobile technology.

The applicability of each mode or style of interaction is determined by context of use; in essence, the various interaction techniques are most powerful and effective when used in combination to create multimodal user interfaces that accord with the contextual requirements of the application and user. There are no hard and fast rules governing how these techniques should be used or combined; innovation is the driving force at present. Mindful of their social acceptability, we need to combine new, imaginative techniques to derive the maximum usability for mobile devices. We need to strive to ensure that users control technology and prevent the complexities of the technology controlling users. We need to eliminate the perception that m-commerce is difficult to use. Most importantly, we need to design future m-interaction so that it is as easy to use as Alexander Graham Bell's old-fashioned telephone—that is, so that users can focus on the semantics of the task they are using the technology to achieve rather than the mechanics of the technology itself.

REFERENCES

Barfield, W. & Caudell, T. (2001). *Fundamentals of wearable computers and augmented reality*. Mahwah, NJ: Lawrence Erlbaum Associates.

Beasly, R., Farley, M., O'Reilly, J., & Squire, L. (2002). *Voice application development with VoiceXML.* SAM Publishing.

Brewster, S.A. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing, 6*(3), 188-205.

Brewster, S.A., Lumsden, J., Bell, M., Hall, M., & Tasker, S. (2003). *Multimodal 'eyes-free' interaction techniques for mobile devices.* Paper presented at the Human Factors in Computing Systems - CHI 2003, Ft Lauderdale, USA.

Cohen, M. & Ludwig, L.F. (1991). Multidimensional audio window management. *International Journal of Man-Machine Studies, 34*(3), 319 - 336.

Fiedlander, N., Schlueter, K., & Mantei, M. (1998). Bullseye! When Fitt's Law doesn't fit. *Paper presented at the ACM CHI'98*, Los Angeles.

Geelhoed, E., Falahee, M., & Latham, K. (2000). Safety and comfort of eyeglass displays. In P. Thomas & H.W. Gelersen (Eds.), *Handheld and ubiquitous computing* (pp. 236-247). Berlin: Springer.

Harrison, B., Fishkin, K., Gujar, A., Mochon, C., & Want, R. (1998). *Squeeze me, hold me, tilt me! An exploration of manipulative user interfaces.* Paper presented at the ACM CHI'98, Los Angeles.

Holland, S. & Morse, D.R. (2001). *Audio GPS: Spatial audio navigation with a minimal attention interface.* Paper presented at the Mobile HCI 2001: Third International Workshop on Human-Computer Interaction with Mobile Devices, Lille, France.

Jarvenpaa, S.L., Lang, K.R., Takeda, Y., & Tuunainen, V.K. (2003). Mobile commerce at crossroads. *Communications of the ACM, 46*(12), 41-44.

Lai, J. & Yankelovich, N. (2000). Conversational speech interfaces. *The human computer interaction handbook* (pp. 698-713). Lawrence Erlbaum Associates Publishers.

Pirhonen, P., Brewster, S.A., & Holguin, C. (2002). *Gestural and audio metaphors as a means of*

control in mobile devices. Paper presented at the ACM-CHI 2002, Minneapolis, MN.

Sarker, S. & Wells, J.D. (2003). Understanding mobile handheld device use and adoption. *Communications of the ACM, 46*(12), 35-40.

Savidis, A., Stephanidis, C., Korte, A., Crispien, K., & Fellbaum, C. (1996). *A generic direct-manipulation 3D-auditory environment for hierarchical navigation in non-visual interaction.* Paper presented at the ACM ASSETS'96, Vancouver, Canada.

Sawhney, N. & Schmandt, C. (2000). Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions on Computer-Human Interaction, 7*(3), 353-383.

Urbaczewski, A., Valacich, J.S., & Jessup, L.M. (2003). Mobile commerce: Opportunities and challenges. *Communications of the ACM, 46*(12), 30-32.

Venkatesh, V., Ramesh, V., & Massey, A.P. (2003). Understanding usability in mobile commerce. *Communications of the ACM, 46*(12), 53-56.

KEY TERMS

Auditory Icon: Icons which use everyday sounds to represent application objects or activities.

Earcon: Abstract, synthetic sounds used in structured combinations whereby the musical qualities of the sounds hold and convey information relative to application objects or activities.

M-Commerce: Mobile access to, and use of, information which, unlike e-commerce, is not necessarily of a transactional nature.

Modality: The pairing of a representational system (or mode) and a physical input or output device.

Mode: The style or nature of the interaction between the user and the computer.

Multimodal: The use of different modalities within a single user interface.

Soundscape: The design of audio cues and their mapping to application objects or user actions.

User Interface: A collection of interaction techniques for input of information/commands to an application as well as all manner of feedback to the user from the system that allow a user to interact with a software application.

Global Navigation Satellite Systems

Phillip Olla
Brunel University, UK

INTRODUCTION

There is a need to determine precise ground locations for use in a variety of innovative and emerging applications such as earth observation, mobile-phone technology, and rescue applications. Location information is pertinent to a large number of remote sensing applications, some of which support strategic tasks such as disaster management, earth monitoring, protecting the environment, management of natural resources, and food production. With the availability of high-resolution images, some applications will require a location precision down to 1 m (Kline, 2004). The global navigation satellite systems (GNSSs) provide signals that can serve this purpose; these signals can be incorporated into a large range of innovative applications with immense benefits for the users (Hollansworth, 1999). Satellite navigation is achieved by using a global network of satellites that transmit radio signals from approximately 11,000 miles in high earth orbit. The technology is accurate enough to pinpoint locations anywhere in the world, 24 hours a day. Positions are provided in latitude, longitude, and altitude. This article provides an overview of the GNSSs in operation along with their uses.

BACKGROUND: WHAT IS GNSS?

There are currently two global systems in operation: the Navigation Satellite Timing and Ranging system (NAVSTAR), commonly referred to as the Global Positioning System (GPS) and owned by the United States of America, and GLONASS (Global´naya Navigatsivannaya Sputnikovaya Sistema) of the Russian Federation. A third system called GALILEO is under development by the European Community (EC) countries. The United States and Russia have offered the international community free use of their respective systems. The business model for GALILEO will be similar to GPS for basic users; however, not all applications will be free as some applications that require a high quality of service will have to be paid for.

GNSS is revolutionizing and revitalizing the way nations operate in space, from guidance systems for the International Space Station's (ISS) return vehicle, to the management tracking and control of communication satellite constellations. Using space-borne GNSS and specialized algorithms, a satellite will soon be capable of self-navigation (Hollansworth, 1999). The underlying technologies of the GNSS infrastructure are very similar, and they have been designed to complement each other even though the initial systems were developed for military purposes. They each consist of three segments: the space segment (the satellites), the ground segment (control and monitoring stations), and the user segment (receiver technology). The GNSS satellites transmit codes generated by atomic clocks, navigation messages, and system-status information, modulated on two carrier frequencies.

The International Civil Aviation Organization (ICAO) and the International Maritime Organization (IMO) have accepted GPS and GLONASS as the core of an international civil capability in satellite navigation. The frequency-spectrum bandwidth, allocated by the International Telecommunications Union (ITU) for GNSS-type applications, is 1,559 1,610 MHz. The unique ITU Aeronautical Radio Navigation Satellite Service allocation provides protection against interference from other sources required by civil aviation, maritime shipping, and other critical safety-of-life applications (Hollansworth, 1999).

CURRENT TRENDS: NAVSTAR GLOBAL POSITIONING SYSTEM

The NAVSTAR GPS was developed by the U.S. Department of Defense (DoD). It consists of a constellation of 24 to 27 satellites in operation at any one time (placed in six orbital planes) orbiting the earth at

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

a high altitude (approximately 10,900 miles). Each plane is inclined 55 degrees relative to the equator. The satellites complete an orbit in approximately 12 hours. The signal from the satellite requires a direct line to GPS receivers and cannot penetrate water, soil, walls, or other obstacles such as trees, buildings, and bridges.

GPS satellites broadcast messages via radio signals. Radio signals travel at the speed of light: 186,000 miles per second (NAVSTAR, 2000). A 3-D position on the earth is calculated from distance measurements (using the travel time of the satellite messages) to three satellites. This requires clocks accurate to within a nanosecond on board the satellites. Since clocks in our GPS receivers are not as accurate, to obtain an accurate 3-D position, a fourth satellite measurement is used to compute the receiver clock-offset errors.

The ultimate accuracy of GPS is determined by the sum of several sources of error. Differential correction is required to reduce the error caused by atmospheric interference. This involves placing a GPS receiver on the ground in a known location acting as a static reference point; this is then utilized to identify errors in the satellite data. An error-correction message is transmitted to any other GPS receivers in the local area to correct their position solutions. This real-time differential correction requires radios to transmit the error-correction messages. Alternatively, postprocessed differential correction can be performed on a computer after the GPS data are collected.

Up until May 1, 2000, the U.S. government scrambled GPS signals for reasons of national security. This intentional signal degradation was called selective availability (SA). Because of SA, the positions computed by a single GPS receiver were in error by up to 100 m. Because of pressure from the civilian GPS user community and other reasons, the government agreed to remove SA.

GLONASS

The fully deployed GLONASS constellation is composed of 24 satellites in three orbital planes whose ascending nodes are 120 degrees apart (Glonass Information, 2003). Each satellite operates in circular 19,100-km orbits at an inclination angle of 64.8 degrees, and each satellite completes an orbit in approximately 11 hours and 15 minutes. The spacing of satellites in orbits is arranged so that a minimum of five satellites is in view to users worldwide. The GLONASS constellation provides continuous and global navigation coverage. Each GLONASS satellite transmits a radio-frequency navigation signal containing a navigation message for users. The first GLONASS satellites were launched into orbit in 1982; the deployment of the full constellation of satellites was completed in 1996, although GLONASS was officially declared operational on September 24, 1993. The system is complementary to the United States' GPS, and both systems share the same principles in the data-transmission and -positioning methods. GLONASS is managed for the Russian Federation government by the Russian Space Forces, and the system is operated by the Coordination Scientific Information Center (KNIT) of the Ministry of Defense of the Russian Federation (SPACE and TECH, 2004)

FUTURE TRENDS: GALILEO

GALILEO is the global navigation satellite system being developed by an initiative launched by the European Union and the European Space Agency (ESA). GALILEO will be fully operable by 2008, however, the signal transmission will start in 2005. This worldwide system will be interoperable with GPS and GLONASS, the two other global satellite navigation systems, providing a highly accurate, guaranteed global positioning service under civilian control. A user will be able to get a position with the same receiver from any of the satellites in any combination. GALILEO will deliver real-time positioning accuracy down to the meter range, which is unprecedented for a publicly available system.

It will guarantee availability of the service under all but the most extreme circumstances and will inform users within seconds of a failure of any satellite. This will make it suitable for applications where safety is crucial, such as running trains, guiding cars, and landing aircraft.

The fully deployed GALILEO system consists of 30 satellites (27 operational plus three active spares) positioned in three circular medium-earth-orbit (MEO) planes at an altitude of 23,616 km above the Earth, and with an inclination of the orbital planes of 56 degrees in reference to the equatorial plane.

GALILEO will provide a global search and rescue (SAR) function similar to the existing operational Cospas-Sarsat system. To do so, each satellite will be equipped with a transponder that is able to transfer the distress signals from the user transmitters to the rescue coordination center, which will then initiate the rescue operation. The system will also provide a signal to the user, informing him or her that the situation has been detected and that help is under way. This feature is new and is considered a major upgrade to the current two systems (DGET, 2004).

Negotiations with U.S. administration are currently focusing on the shared use of certain frequency bands, which will allow a combined GPS and GALILEO receiver that will be capable of computing signals from both constellations. This will provide for the best possible performance, accuracy, and reliability. However, since GALILEO will not be available before 2008, current GPS receivers will not be able to receive GALILEO signals.

The critical issue with the current implementation of GNSS for nonmilitary purposes is that some applications require the system to have special features. These features include service guarantee, liability of the service operator, traceability of past performance, operation transparency, certification, and competitive service performance in terms of accuracy and availability. These features do not currently exist in the current systems. New applications are appearing everyday in this huge market, which is projected to reach at least 1,750 million users in 2010 and 3,600 million in 2020 (ESA, 2004).

BUSINESS APPLICATIONS OF GNSS

Benefits to user applications detailed by the ESA (2004) are described below. The anticipated benefit to aviation and shipping operators alone is put at EUR 15 billion between 2008 and 2020. This includes savings generated by more direct aircraft flights through better air-traffic management, more efficient ground control, fewer flight delays, and a single global, multipurpose navigation system.

Future research will also incorporate satellite signals into driving systems. At present, road accidents generate social and economic costs corresponding to 1.5 to 2.5% of the gross national product (GNP) of the European Union. Road congestion entails additional estimated costs of around 2% of the European GNP. A significant reduction in these figures will have considerable socioeconomic benefits; this is additional to the number of lives saved. Vehicle manufactures now provide navigation units that combine satellite location and road data to avoid traffic jams and reduce travel time, fuel consumption, and therefore pollution. Road and rail transport operators will be able to monitor the goods' movements more efficiently, and combat theft and fraud more effectively. Taxi companies now use these systems to offer a faster and more reliable service to customers.

Incorporating the GNSS signal into emergency-services applications creates a valuable tool for the emergency services (fire brigade, police, paramedics, sea and mountain rescue), allowing them to respond more rapidly to those in danger. There is also potential for the signal to be used to guide the blind (Benedicto, Dinwiddy, Gatti, Lucas, & Lugert, 2000); monitor Alzheimer's sufferers with memory loss; and guide explorers, hikers, and sailing enthusiasts.

Surveying systems incorporating GNSS signals will be used as tools for urban development. They can be incorporated into geographical information systems for the efficient management of agricultural land and for aiding environmental protection; this is a critical role of paramount importance to assist developing nations in preserving natural resources and expanding their international trade. Another key application is the integration of third-generation mobile phones with Internet-linked applications (Muratore, 2001). It will facilitate the interconnection of telecommunications, electricity, and banking networks and systems via the extreme precision of its atomic clocks.

CONCLUSION

The role played by the current global navigation satellite systems in our everyday lives is set to grow considerably with new demands for more accurate information along with integration into more applications. The real impact of satellite global positioning on society and industrial development will become evident when GALILEO becomes operational and innovative application outside the arena of transportation and guidance become available.

Some analysts regard satellite radionavigation as an invention that is as significant in its own way as that of the watch: No one nowadays can ignore the time of day, and in the future, no one will be able to do without knowing their precise location (DGET, 2004).

The vast majority of satellite navigation applications are currently based on GPS performances, and great technological effort is spent to integrate satellite-derived information with a number of other techniques in order to reach better positioning precision with improved reliability.

This scenario will significantly change in the short-term future. European regional augmentation of GPS service will start in 2004. Four years later, the global satellite navigation system infrastructure will double with the advent of GALILEO. The availability of two or more constellations will double the total number of available satellites in the sky, therefore enhancing the quality of the services and increasing the number of potential users and applications (DGET, 2004).

REFERENCES

Benedicto, J., Dinwiddy, S. E., Gatti, G., Lucas, R., & Lugert, M. (2000). *GALILEO: Satellite system design and technology developments.* European Space Agency.

DGET. (2004). *GALILEO: European satellite navigation system.* Directorate of General Energy and Technology. Retrieved from *http://europa.eu.int/comm/dgs/energy_transport/galileo/index_en.htm*

ESA. (2004). *Galileo: The European programme for global navigation services.* Retrieved from *http://www.esa.int/esaNA/index.html*

Glonass information. (2003). Retrieved from *http://www.glonass-center.ru/constel.html*

Hollansworth, J. E. (1999). Global Navigation Satellite System (GNSS): What is it? *Space Communications Technology, 2*(1).

Kline, R. (2004). Satellite navigation in the 21st century serving the user better? *Acta Astronautica, 54*(11-12), 937.

Muratore, F. (2001). *UMTS mobile communication of the future*. Chicester, UK: Wiley.

NAVSTAR. (2000). *NAVSTAR Global Positioning System (GPS) facts*. Montana State University. Retrieved 2004 from *http://www.montana.edu/places/gps/*

SPACE and TECH. (2004). Retrieved *http://www.spaceandtech.com/spacedata/constellations/glonass_consum.shtml*

KEY TERMS

Differential Correction: The effects of atmospheric and other GPS errors can be reduced using a procedure called differential correction. Differential correction uses a second GPS receiver at a known location to act as a static reference point. The accuracy of differentially corrected GPS positions can be from a few millimeters to about 5 m, depending on the equipment, time of observation, and software processing techniques.

Geostationary Satellite (GEO): A geostationary satellite orbits the earth directly over the equator, approximately 22,000 miles up. At this altitude, one complete trip around the earth (relative to the sun) takes 24 hours. The satellite remains over the same spot on the earth's surface at all times and stays fixed in the sky at any point from which it can be seen from the surface. Weather satellites are usually of this type. Satellites, spaced at equal intervals (120 angular degrees apart), can provide coverage of the entire civilized world. A geostationary satellite can be accessed using a dish antenna aimed at the spot in the sky where the satellite hovers (http://whatis.techtarget.com/).

Low Earth Orbit (LEO): This satellite system employs a large fleet of "birds," each in a circular orbit at a constant altitude of a few hundred miles. The orbits take the satellites over, or nearly over, the geographic poles. Each revolution takes approximately 90 minutes to a few hours. The fleet is arranged in such a way that, from any point on the surface at anytime, at least one satellite is in line of sight. A well-designed LEO system makes it possible for anyone to access the Internet via a wireless device from any point on the planet (http://whatis.techtarget.com/).

Satellite: A satellite is a specialized wireless receiver and transmitter that is launched by a rocket and placed in orbit around the earth. There are hundreds of satellites currently in operation. They are used for such diverse purposes as weather forecasting, television broadcasting, amateur radio communications, Internet communications, and the Global Positioning System (http://whatis.techtarget.com/).

Satellite Constellation: A group of satellites working in concert is known as a satellite constellation. Such a constellation can be considered to be a number of satellites with coordinated coverage, operating together under shared control, and synchronised

Going Virtual

Evangelia Baralou
University of Sterling, Scotland

Jill Shepherd
Simon Fraser University, Canada

WHAT IS VIRTUALITY AND WHY DOES IT MATTER?

Virtuality is a socially constructed reality mediated by electronic media (Morse, 1998). Characterized by the dimension of time-space distantiation (Giddens, 1991), virtuality has an impact on the nature and dynamics of knowledge creation (Thompson, 1995). The relentless advancement of Information and Communication Technology (ICT) in terms both of new technology and the convergence of technology (e.g., multimedia) is making virtual networking the norm rather than the exception. Socially, virtual communities are more dispersed, have different power dynamics, are less hierarchical, tend to be shaped around special interests, and are open to multiple interpretations, when compared to face-to-face equivalents. To successfully manage virtual communities these differences need firstly to be understood, secondly the understanding related to varying organizational aims and thirdly, the contextualised understanding needs to be translated into appropriate managerial implications.

In business terms, virtuality exists in the form of life style choices (home-working), ways of working (global product development teams), new products (virtual themeparks), and new business models (e.g., Internet dating agencies). Socially, virtuality can take the form of talking to intelligent agents, combining reality and virtuality in surgery (e.g., using 3D imaging before and during an operation), or in policy making (e.g., combining research and engineering reports with real satellite images of a landscape with digital animations of being within that landscape, to aid environmental policy decisions).

Defining virtuality today is easy in comparison with defining, understanding and managing it on an ongoing basis. As the title "going virtual" suggests, virtuality is a matter of a phenomenon in the making, as we enter into it during our everyday lives, as the technology develops and as society changes as a result of virtual existences. The relentless advances in the technical complexity which underlies virtual functionality and the speeding up and broadening of our lives as a consequence of virtuality, make for little time and inclination to reflect upon the exact nature and effect of going virtual. As it pervades the way we live, work and play at such a fast rate, we rarely have the time to stop and think about the implications of the phenomenon.

The aim of what follows is therefore to reflexively generate an understanding of the techno-social nature of virtuality on the basis that such an understanding is a prerequisite to becoming more responsible for its nature and effects. Ways of looking at virtuality are followed by some thoughts on the managerial implications of "going virtual".

A TECHNO-SOCIAL VIEW OF VIRTUALITY

Marx foresaw how the power of technological innovation would drive social change and how it would influence and become influenced by the social structure of society and human behaviour (Wallace, 1999). This interrelationship means that an understanding of virtuality needs to start from the theoretical acceptance of virtuality as a social reality; considering it involves human interaction associated with digital media and language in a socially constructed world (Morse, 1998). More specifically, Van Dijk (1999) suggests that going virtual, in comparison with face to face interaction, is characterised by:

- A less stable and concrete reality without time, place and physical ties

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

- More abstract interaction which affects knowledge creation
- A networked reality which both disperses and concentrates power, offering new ways of exercising power
- Diffused and less hierarchical communities and interaction due to the more dynamic flow of knowledge and greater equality in participation
- A reality often shaped around special interests

Each of these areas is explored below, with the aim of drawing out the issues such that the managerial implications can be discussed in the following section. The emphasis is not on the technology, but on the socio-managerial implications of how the technology promotes and moulds social existence within virtual situations.

A REALITY WHICH IS LESS STABLE AND CONCRETE

Arguably, the most fundamental characteristic of virtuality is the first on this list, namely time-space distantiation (Giddens, 1991). Prior to the development of ICTs, the main mode of communication between individuals was face-to-face interaction in a shared place and time. The presence of a shared context during face to face contact provides a richness, allowing for the capacity to interrupt, repair, feedback and learn, which some see as an advantage (Nohria & Eccles, 1992, cited by Metiu & Kogut, 2001). In a virtual context, individuals interact at a distance and can interact asynchronously in cyberspace through the mediation of ICTs. The absence of shared context and time has an impact on communication (Metiu & Kogut, 2001; Thompson, 1995).

A MORE ABSTRACT REALITY

In virtuality, a narrowed range of nonverbal symbolic cues can be transmitted to distant others (Foster & Meech, 1995; Sapsed, Bessant, Partington, Tranfield, & Young, 2002; Wallace, 1999), albeit technology advancement is broadening the spectrum. Social cues associated with face-to-face co-presence are deprived, while other symbolic cues (i.e., those linked to writing) are accentuated (Thompson, 1995). The additional meaning found in direct auditory and visual communication, carried by inflections in the voice tone, gestures, dress, posture, as well as the reflexive monitoring of others' responses, is missing. Human senses such as touch, smell, taste cannot be stimulated (Christou & Parker, 1995). Virtuality is a more abstract form of reality. These symbolic cues convey information regarding the meaning individuals assign to the language they use, as well as the image they want to project while expressing themselves. In this sense man first went virtual when language evolved, given language was arguably the first abstract space man inhabited.

Understanding the social impact of mediated interaction is helped by thinking in terms of the spaces within which individuals interact (Goffman 1959, cited by Thompson, 1995). A distinction is made between individuals interacting within and between easily accessible front regions, separated in space and perhaps in time from their respective back regions into which it is difficult, if not impossible, to intrude.

In a face-to-face context, social interaction takes place in a shared front region, a setting that stays put geographically speaking (e.g., an office, a class), which can be directly observed by others and is related to the image the individual wants to project. Actions that seem to be inappropriate or contradictory, for that image, are suppressed and reserved in the back region, for future use. It is not always easy to identify the distinction between the front region and the back region, as there can be regions which function at one time and in one sense as a front region and at another time and in another sense as a back region. For example, a manager in his office with clients or other employees can be considered as acting in a front region, whereas the same geographical setting can be thought of as the back region before or after the meeting.

In virtuality, the separation of back and front regions can lead to a loss of the sense of normal social presence as individuals become disembodied beings that can potentially be anywhere in the universe without the actual embodied presence (Dreyfus, 2001). Reality appears anonymous, opaque and inaccessible, without the sociability, warmth, stability and sensitivity of face-to-face communication (Short, Williams, & Christie, 1976; van Dijk, 1999). The dichotomy between appearance and reality set up by Plato is intensified. People operating virtually spend

more time in an imaginary virtual world than in the real world (Woolgar, 2002).

That said, such disembodied social presence creates opportunities. Whilst interacting in a virtual as well as in a face-to-face context, participants construct their own subjective reality, using their particular experience and life history; and incorporating it into their own understanding of themselves and others (Duarte & Snyder, 1999). In a virtual context, individuals live in each other's brains, as voices, images, or words on screens, which arguably makes them become capable of constructing multiple realities, of trying out different versions of self, to discover, what is "me" and what is "not me", versions of which they are in greater control, taking also with them the reality, or indeed the realities they are familiar with (Turkle, 1995; Whitty, 2003).

Individuals can thus take advantage of the lack of context by manipulating front and back regions, more consciously inducing and switching to multiple personas, projecting the image they want in the cyberspace, thus controlling the development of their social identity, based on the different degrees of immersiveness (Morse, 1998; van Dijk, 1999). From this point of view, it can be argued that the virtual context empowers individuals. Interestingly, multimedia provides more "natural" interaction allowing, for example, the use of voice through Internet telephony and the bringing back into the social frame, for example, of body language and dress, through Web-cams. Does therefore the advancement of ICT mean that virtuality will become more "normal" or will the habit of self-identity construction within virtual reality remain?

POWER AND EMPOWERMENT

Giddens (1991) suggests that virtuality offers new modes of exercising power and that virtuality is creating a more reflective society due to the massive information received. This can be questioned on the basis that more is read than written and more is listened to than spoken within the virtual world, which could shape an increasingly passive society hijacked by its own knowledge drifting around the infinite and complex reality of cyberspace. The relationship between power and knowledge in a virtual context remains under researched. Perhaps it is knowledge itself, which becomes more powerful. It has been found (Franks, 1998), that in an organizational virtual context, the demands of quick changes in knowledge requirements result in managers not being able to keep up. They entrust related decision making to the remote employees. Although, this empowerment enhances greater equality in participation, the property rights of the produced knowledge remain organizational, which can make individuals feel weaker and objects of control and pervasiveness given that their whole online life can ironically also be remotely supervised and archived (Franks, 1998; Ridings, Gefen, & Arinze, 2002). Power dynamics are therefore usually different in virtual reality when compared to face-to-face reality.

A REALITY OFTEN SHAPED AROUND SPECIAL INTERESTS

The claim that virtuality shapes communities around shared interests can be understood in relation to the way traditional relationships are shaped and maintained in a virtual context. Dreyfus (2001) emphasizes the withdrawal of people from traditional relations, arguing that the price of loss of the sense of context in virtuality is the inability to establish and maintain trust within a virtual context (Giddens, 1991). Trust has been in the centre of studies on human relations (Handy, 1995). Hosmer (1995, p. 399) defines trust as the "expectation by one person, group, or firm of ethical behaviour on the part of the other person, group, or firm in a joint endeavour or economic exchange". Traditionally, individuals establish their relations based on trust and interact inside a context of social presence, which is affected in virtuality by the physical and psychological distance, by loose affiliations of people that can fall apart at any moment, by a lack of shared experiences and a lack of knowledge of each other's identity.

Sapsed et al. (2002) suggest that trust in a virtual environment is influenced by the accessibility, reliability and compatibility in ICT, is built upon shared interests and is maintained by open and continuous communication. The quantity of information shared, especially personal information, is positively related to trust (Jarvenpaa & Leidner, 1999; Ridings et al., 2002). Being and becoming within virtual communities also depends more on cognitive elements (e.g.,

competence, reliability, professionalism) than affective elements (e.g., caring, emotional, connection to each other), as emotions cannot be transmitted that easily (Meyerson et al., 1996, cited by Kanawattanachai & Yoo, 2002).

That said, virtual communities do exist which are perhaps breaking with tradition. Consequently, what is "normal" or "traditional" in time is likely to change. In such communities the lack of trust allows views to be expressed more openly, without emotion and people are more able to wander in and out of communities. Special interests are more catered for as minority views can be shared. An absence of trust is less of an issue. The Net is always there and can be more supportive than a local community, making the Net more real than reality and more trustworthy.

In considering the above characteristics of virtuality within management, three factors of organizational life are taken into consideration in the following section; the first is context, the second is the organising challenges which emerge within organizational contexts, the third is the matter of taking into account advancements in technology.

THE MANAGERIAL IMPLICATIONS OF GOING VIRTUAL

Organizationally the ability to communicate virtually brings increased productivity and opportunity. Getting more out of going virtual requires placing an in-depth understanding of virtuality along side organizational context in terms of organizational aims as well as managing the tensions which arise out of those unique organizational contexts. It involves constantly appraising ICT technology convergence and advancement to establish and re-establish what virtuality and virtual networking mean.

The first managerial implication is that managers must be aware of the nature of virtuality in terms of the strategic intent of the organization. Beyond increases in operational productivity, strategically must the more abstract interaction of the virtual world be countered or is it an enabler of the aim? More specifically, if an organization wishes to share knowledge without any variation or interpretation, meaning without any knowledge creation, within a confined community, then going virtual can be problematic, especially if counter measures are not taken to reduce the chances of knowledge creation, community boundaries being broken or made impenetrable, and sharing being reduced through a lack of trust. Thus, e-mails can be sent to the wrong group of people, the content of chat-rooms can be far more risqué than would be the case face to face, interest groups can self-organise to lobby against convention, people can appear to be other than they really are, and/or the message can be misinterpreted with negative outcomes.

However, where knowledge creation is desired, then these supposed disadvantages can be turned into advantages. For example, a lack of physical shared context within virtual environments can create a way of sharing knowledge that is tacit, abstract, difficult to describe but which can also be a source of core competence. Engineers working within CADCAM systems across organizational sites is an example of how going virtual can create a way of communicating unique to that community and difficult to imitate. The challenge here is to understand the way of creating and sharing knowledge and how it might be preserved.

Equally, intranet chat rooms aimed at sharing of ideas can be more innovative because of the lack of social context. In this sense going virtual allows managers to; take risks they would not do in face to face settings, more easily misinterpret others to create new knowledge, not allow sources of bias present in face to face encounters to creep into knowledge sharing and creation, participate in conversations which they might otherwise not participate in because they are shy or do not know that the conversation is taking place because the conversation is within strict boundaries. Thus, the anonymous, self-organizing characteristics of going virtual can be advantageous.

On important question remains: as technology becomes more advanced, converging to bring the use of all senses into the virtual realm and as it pervades our everyday lives, will virtuality become as real, as normal, as common as physicality? Virtuality exists in the making as individuals and technologies co-evolve. Indeed, "real virtuality" is talked of, in which within a virtual setting, reality (that is people's material and symbolic nature) is captured and exchanged. Perhaps real virtuality is not a channel through which to experience a more abstract, networked life—it is life, it is the experience.

So to conclude, organizations must be aware of whether the aim within the virtual space is to reconsider reality and to create knowledge or to communi-

cate reality with no creation of knowledge. In either case the moderating role of power in knowledge flows, the manipulation of front and back regions, as well as the dynamics and nature of community membership must be appreciated and managed appropriately.

Finally as part of our social fabric, virtuality is becoming more natural and more traditional in the sense we are becoming more accustomed to the role it plays in our lives, the technology that underpins it, the opportunities it brings. Perhaps "going virtual" above all involves accepting that virtuality is as real as reality, but needs to be equally managed based on in-depth understanding and reflexive practice.

REFERENCES

Christou, C. & Parker, A. (1995). Visual realism and virtual reality: A psychological perspective. In Carr, K. & England, R. (Eds.), *Simulated and virtual realities: Elements of perception*. USA: Taylor and Francis.

Dreyfus, H. (2001). *On the Internet*. London: Routledge.

Duarte, D. & Snyder, N. (1999). M*astering virtual teams: Strategies, tools, and techniques that succeed*. CA: Jossey-Bass Publishers.

Foster, D. & Meech, J. (1995). Social dimensions of virtual reality. In K. Carr & R. England (Eds.), *Simulated and virtual realities: Elements of perception*. USA: Taylor & Francis.

Giddens, A. (1991). *The consequences of modernity.* CA: Stanford University Press.

Goffman, E. (1959). *The presentation of self in everyday life.* New York: Doubleday Anchor.

Handy, C. (1995). Trust and the virtual organization. *Harvard Business Review,* May-June, 40-50.

Hosmer, L. (1995). Trust: The connection link between organizational theory and philosophical ethics. *Academy of Management Review, 20*, 379-403.

Jarvenpaa, S.L. & Leidner, D.E. (1999). Communication and trust in global virtual teams. *Organization Science, 10*, 791-815.

Kanawattanachai, P. & Yoo, Y. (2002). Dynamic nature of trust in virtual teams. *Journal of Strategic Information Systems, 11*, 187-213.

Metiu, A. & Kogut, B. (2001). *Distributed knowledge and the global organization of software development.* Working paper.

Morse, M. (1998). *Virtualities: Television, media art, and cyberculture.* USA: Indiana University Press.

Ridings, C.M., Gefen, D., & Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. *Journal of Strategic Information Systems, 11*, 271-295.

Sapsed, J., Bessant, J., Partington, D. Tranfield, D., & Young, M. (2002). Teamworking and knowledge management: A review of converging themes. *International Journal of Management Reviews, 4*(1), 71-85.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications.* New York: Wiley.

Thompson, J. (1995). *The media and modernity: A social theory of the media.* UK: Polity Press.

Turkle, S. (1995). *Life on the screen*. London: Weidenfeld and Nicholson.

van Dijk, J. (1999). *The network society*. London: SAGE.

Wallace, P. (1999). *The psychology of the Internet*. USA: Cambridge University Press.

Whitty, M. (2003). Cyber-flirting: Playing at love on the Internet. *Theory & Psychology, 13*(3), 339-357.

Woolgar, S. (2002). *Virtual society? Technology, cyberbole, reality*. UK: Oxford University Press.

KEY TERMS

Electronic Media: Interactive digital technologies used in business, publishing, entertainment, and arts.

Front and Back Region: Front region is a setting that stays put geographically speaking, (e.g., an of-

fice, a class). Back region is a setting which cannot be easily intruded upon.

Knowledge: An individual and social construction that allows us to relate to the world and each other

Mediated Interaction: Involves the sender of a message being separated in time and space from the recipient.

Reflective Society: One that takes a critical stance to information received and beliefs held.

Social-Construction: Anything that could not have existed had we not built it (Boghossian, 2001, available at http://www.douglashospital.qc.ca/fdg/kjf/38-TABOG.htm).

Social Realities: Constructs which involve using the same rules to derive the same information (individual beliefs) from observations (Bittner, S., available at www.geoinfo.tuwien.ac.at/projects/revigis/carnuntum/Bittner.ppt).

Virtuality: A socially constructed reality, mediated by electronic media.

Heterogeneous Wireless Networks Using a Wireless ATM Platform

Spiros Louvros
COSMOTE S.A., Greece

Dimitrios Karaboulas
University of Patras, Greece

Athanassios C. Iossifides
COSMOTE S.A., Greece

Stavros A. Kotsopoulos
University of Patras, Greece

INTRODUCTION

Within the last two decades, the world of telecommunications has started to change at a rapid pace. Data traffic, where the information is transmitted in the form of packets and the flow of information is bursty rather than constant, now accounts for almost 40 to 60% of the traffic that is transmitted over the backbone telecommunication networks (Esmailzadeh, Nakagawa, & Jones, 2003). In addition to data traffic, video traffic (variable rate with real-time constraints) was made possible by low-cost video-digitizing equipment (Houssos et al., 2003).

Asynchronous transfer mode (ATM) technology is proposed by the telecommunications industry to accommodate multiple traffic types in a very high-speed wireline-backbone network. Briefly, ATM is based on very fast (on the order of 2.5 Gbits/sec or higher; Q.2931 ATM network signaling specification, ITU, n.d.) packet-switching technology with 53-byte-long packets called cells being transmitted through wireline networks running usually on fiber-optical equipment.

Wireless telecommunications networks have broken the tether in wireline networks and allow users to be mobile and still maintain connectivity to their offices, homes, and so forth (Cox, 1995). The wireless networks are growing at a very rapid pace; GSM-based (global system mobile) cellular phones have been successfully deployed in Europe, Asia, Australia, and North America (Siegmund, Redl, Weber, & Oliphant, 1995). For higher bit-rate wireless access, the *Universal Mobile Telecommunications System* (UMTS) has been already developed. Finally, for heterogeneous networks, including ex-military networks, ad hoc cellular and *high altitude stratospheric platform* (HASP) technologies are under development, and standardization for commercial data transmissions in heterogeneous environments has launched.

A *wireless ATM transmission network* provides a natural wireless counterpart to the development of ATM-based wireline transmission networks by providing full support for multiple traffic types including voice and data traffic in a wireless environment. In this article, an architecture for a wireless ATM transmission platform is presented as a candidate for the interconnection of heterogeneous, wireless cellular networks.

TECHNICAL BACKGROUND

Wireless Mobile Network Overview

In 1991 the European Telecommunication and Standardization Institute (ETSI) accepted the standards for an upcoming mobile, fully digital and cellular communication network: GSM. It was the first Pan-European mobile telephone-network standard that replaced all the existing analogue ones.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Broadband integrated-services data networks (B-ISDNs) are the state-of-the-art technology in today's wired telecommunication links. The main feature of the B-ISDN concept is the support of a wide range of voice and nonvoice applications in the same network. Mobile networks have to follow the evolution of fixed networks in order to provide moving subscribers with all the services and applications of fixed subscribers. The result of this effort (although somewhat restrictive in terms of realizable bit rates) was another evolution in mobile networks: general packet radio services (GPRSs) and the enhanced data for GSM evolution (EDGE) network (usually referred to as 2.5G), with rates of up to 115 Kb/s and 384 Kb/s, respectively, when fully exploited.

UMTS is the realization of a new generation of telecommunications technology for a world in which personal services will be based on a combination of fixed and mobile services to form a seamless end-to-end service for the subscriber. Generally speaking, UMTS follows the demand posed by moving subscribers of upgrading the existing mobile cellular networks (GSM, GPRS) in nonhomogeneous environments.

3.5G and 4G systems (Esmailzadeh et al., 2003) are already under investigation. Aiming to offer "context-aware personalized ubiquitous multimedia services"(Houssos et al., 2003), 3.5G systems promise rates of up to 10 Mb/s (3GPP [3rd Generation Partnership Project] Release 5), while the use of greater bandwidth may raise these rates even more in 4G (Esmailzadeh et al.). On the other hand, in the last five years a standardization effort has started for the evolution of WLANs (wireless local-area networks) in order to support higher bit rates in hot spots or business and factory environments with a cell radius in the order of 100 m. For example, IEEE 802.11 variants face rates of up to 11 Mb/s (802.11b) and 54 Mb/s (802.11a/g), while rates in excess of 100 Mb/s have already been referred (Simoens, Pellati, Gosteau, Gosse, & Ware, 2003). European HIPERLAN/2 supports somewhat lower rates but with greater cell coverage and enhanced MAC (medium access control) protocols. In any case, 4G and WLAN technology are going to be based on an IP (Internet protocol) backbone between APs and access controllers, or routers and the Internet. Mobile IPv4 and IPv6 are already under investigation (Lach, Janneteau, & Petrescu, 2003) to provide user mobility support for context-type services.

Heterogeneous Wireless Networks Overview

In the near future, the offered communication services to mobile users will be supported by combined heterogeneous wireless networks. This situation demands actions in the following engineering issues.

- Integration with existing technologies in the radio network and in the switching levels of the involved combined wireless communication networks.
- Reengineering of the appropriate interface units at the link layers of the involved networks in order to support optimum access procedures to the corresponding media.
- Implementation of systemic handover procedures in order to combine the independent handover and roaming procedures of the involved wireless networks.
- Introduction of new methods and techniques to provide a number of effective security measures.
- Introduction of advanced ATM procedures in order to support optimum information routing between the main nodes of the combined wireless network.
- New protocol versions of the existing technologies in order to support interoperability demands.

It is worthwhile to mention that the possible involved wireless networks that are going to set the futuristic heterogeneous environment belong to the following categories.

- WLANs covering small geographical areas. In this case the WLANs with the adopted protocols IEEE 802.11a and IEEE 802.11g, and supporting user services on the orthogonal frequency-division multiple-access (OFDM) technique seem to appear as the great scientific interest (Simoens et al., 2003).
- Ad hoc networks, operating in specific geographical areas using the IEEE 802.11b protocol, will be involved on nested schemes under the technology of the existing cellular communication systems.

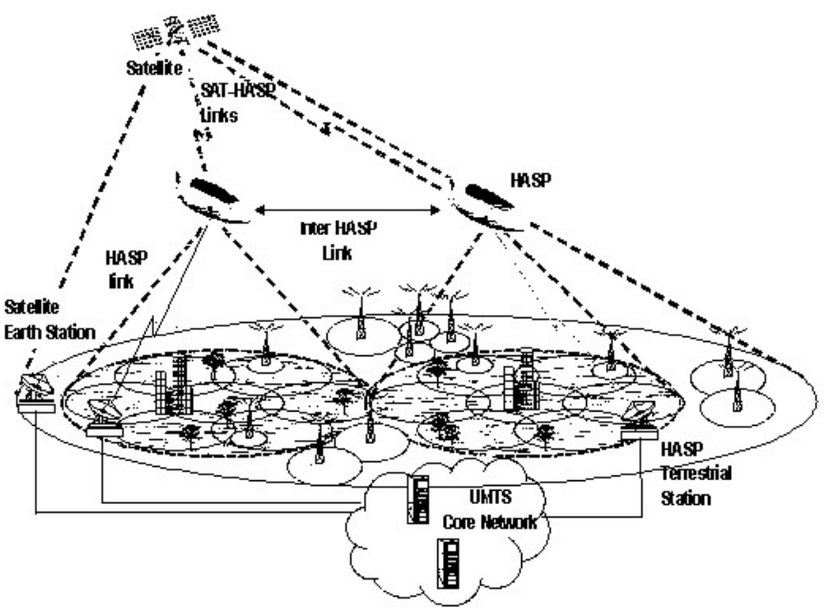
- Cellular mobile networks of 2.5G (i.e., GPRS) and 3G (wideband CDMA [code division multiple access]) will cover geographical areas with mixed cell sizes (i.e., pico-, micro-, and macrocell). In this case, cellular-aided mobile ad hoc networking becomes a very interesting and "hot" research area for reaching the heterogeneous combination of the involved two different types of wireless networks.
- High-altitude stratosphere platforms will soon cover the non-line-of-sight communication applications and are going to support satellite-like communications with the advantage of small energy demands on the used portable and mobile phones. The SkyStation, SkyNet, SkyTower, and EuroSkyWay projects declare new promises to the applications for a large-scale geographical coverage (Varquez-Castro, Perez-Fontan, & Arbesser-Rastburg, 2002).
- Satellite communications networks using low earth-orbiting (LEO) and medium earth-orbiting (MEO) satellites will continue to offer their communication services and to expand the communication activities of the terrestrial wireless communication networks.

The futuristic technology convergence on the heterogeneous wireless networking environment is depicted in Figure 1. The lower layers consist of the land mobile networks (GSM, UMTS, WLAN, general ad hoc networks). Above these layers exist the HASP platform either as an overlay umbrella cell or as an overlay switching and interconnecting platform among the different switching protocols of the lower layers. Finally, on the top of all is the high-altitude satellite network.

ATM Overview

ATM technology is proposed by the telecommunications industry to accommodate multiple traffic types in a very high-speed wireline network. The basic idea behind ATM is to transmit all information in small, fixed-size packets called *ATM cells* over all transmission channels (wired or wireless). Having fixed-size packets of information for transmission can emulate the circuit-switching technique of traditional telephony networks and at the same time take advantage of the best utilization of the transmission-line bandwidth. Hence, it operates asynchronously and it can continuously switch information from and to different networks (voice, video, data) with variable bit rates. The responsible nodes for asynchronous operation are called *ATM switches*. They consist of interfaces in order to communicate with various heterogeneous networks such as LANs (local-area networks), WANs (wide-area networks), and so forth. All these networks transmit

Figure 1. Wireless networking technology convergence



information in different bit rates, and the ATM switches (through the ATM layer of B-ISDN or IP hierarchy) divide this heterogeneous information (using special ATM adaptation layers in terms of OSI [open systems interconnection] layer structure) into fixed-size packets of 48 bytes to accommodate them into the ATM cells.

ATM supports a QoS (quality of service) concept, which is a mechanism for allocating resources based upon the needs of the specific application. The ATM Forum (1996; Rec. TM 4.0) has defined the corresponding service categories (constant bit rate [CBR] for real-time applications, such as videoconferences with strict QoS demands; real-time variable bit rate [rt-VBR] for bursty applications such as compressed video or packetised voice; and so forth).

CHALLENGES IN WIRELESS ATM NETWORKS

The ATM-network architecture has to be redesigned to support wireless users. The use of wireless ATM networks as an interconnection medium among several wireless platforms in a heterogeneous environment is important. So far, WLANs using wireless Ethernet and wireless ATMs have been considered during the evolution toward 4G and beyond-4G wireless mobile heterogeneous networks (Figure 3). Supporting wireless users presents two sets of challenges to the ATM network. The first set includes problems that arise due to the mobility of the wireless users. The second set is related to the provisioning of access to the wireless ATM network.

Mobility of Wireless Users

The ATM standards proposed by the International Telecommunications Union (ITU) are designed to support wireline users at fixed locations (Lach et al., 2003); on the other hand, wireless users are mobile. Current ATM standards do not provide any provisions for the support of location lookup and registration transactions that are required by mobile users (Lach et al.). They also do not support handoff and rerouting functions that are required to remain connected to the backbone ATM network during a move.

The user identification (UID) numbers in wireline networks may be used for the routing of connections to the user; in contrast, the identification number for a wireless user may only be used as a key to retrieve the current location information for that user. The location information for wireless users is usually stored in a database structure that is distributed across the network (Jain, Rajagopalan, & Chang, 1999; Rajagopalan, 1995; Simoens et al., 2003). This database is updated by registration transactions that occur as wireless users move within the wireless network. During a *connection setup*, the network database is used to locate and route connections to the user.

If a wireless user moves while he or she is communicating with another user or a server in the network, the network may need to transfer the radio link of the user between radio access points in order to provide seamless connectivity to the user. The transfer of a user's radio link is referred to as handoff. In this article, *mobility signaling protocols*, designed to implement mobility-related functionality in an ATM network, are described.

Providing Access to the Wireless ATM Network

A key benefit of a wireless network is providing tetherless access to the subscribers. The most common method for providing tetherless access to a network is through the use of radio frequencies. There are two problems that need to be addressed while providing access to an ATM network by means of radio frequencies.

- **Error Performance of the Radio Link:** ATM networks are designed to utilize highly reliable fiber-optical or very reliable copper-based physical media. ATM does not include error correction or checking for the user-information portion of an ATM packet. In order to support ATM traffic in a wireless ATM network, the quality of the radio links needs to be improved through the use of equalization, diversity, and error correction and detection to a level that is closer to wireline networks. There are a number of solutions that combine these techniques to improve the error performance of wireless networks. Some of these solutions may be

Figure 2. Components of wireless ATM architecture

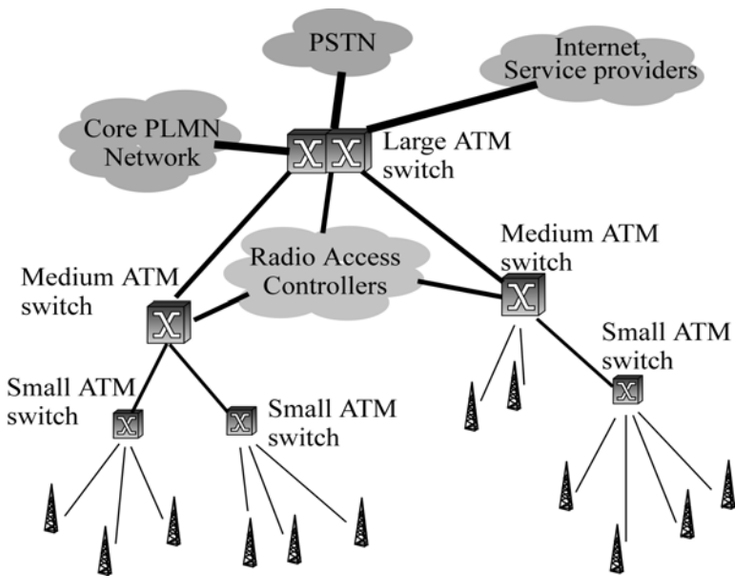
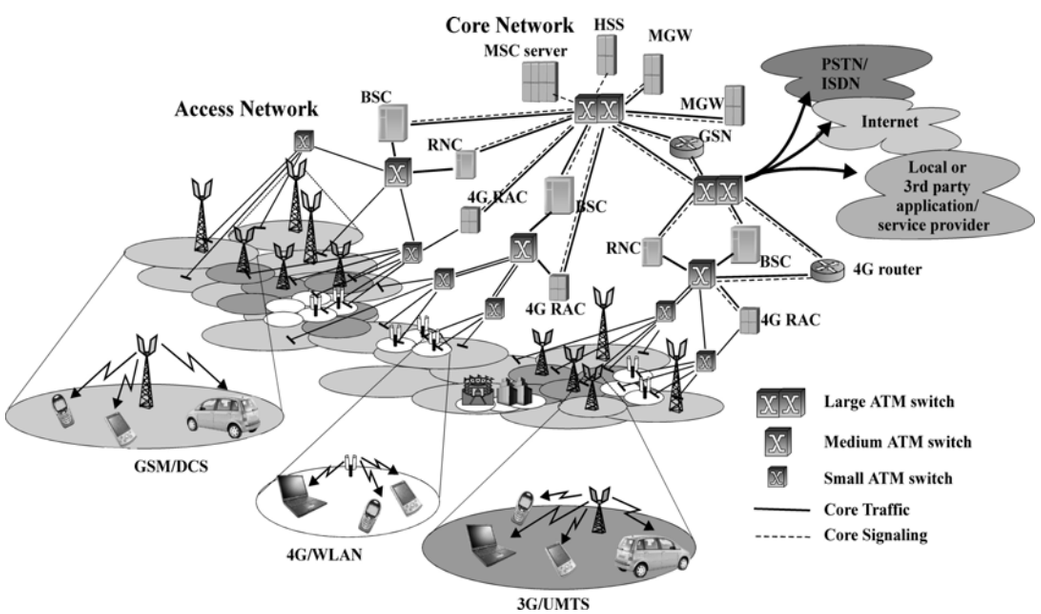


Figure 3. Future heterogeneous mobile network architecture with different technologies (2G/3G/4G) engaging a multi-layer wireless ATM interconnection architecture



found in Acampora (1994), ATM Forum (1996), Chan, Chan, Ko, Yeung, and Wong (2000), and Cox (1991, 1995).

- **Medium Access for Wireless ATM Networks:** A wireless ATM network needs to support multiple traffic types with different priorities and quality-of-service guarantees. In contrast to the fiber-optical media in wireline networks, radio bandwidth is a very precious resource for the wireless ATM network. A medium-access control protocol that supports multiple users, multiple connections per user, and service priorities with quality-of-service requirements must be developed in order to maintain full compatibility with the existing ATM protocols. This medium-access protocol needs to make maximum use of the shared radio resources and needs to achieve full utilization of the radio frequencies in a variety of environments.

WIRELESS ATM-PLATFORM DESCRIPTION

This section introduces our wireless ATM-network architecture. It describes the components of the wireless ATM network and the functions of these components. It also describes the registration- (location) area concept.

Components of the Wireless ATM Network

A wireless ATM-network architecture is based on the registration-area concept. A registration area consists of radio ports, radio-port controllers (medium and small ATM switches), possibly a database, and the physical links that interconnect the parts of the registration area (Figure 2).

The wireless ATM network is designed as a microcellular network for the reasons described in Cox (1991), and Wang and Lee (2001). The typical coverage of a radio port in a microcellular network varies between 0.5 km to 1 km (Cox); therefore, a fairly large number of radio ports are required in order to maintain full coverage of a given geographical area. Consequently, the radio ports in a microcellular network must be economical radio modems that are small enough to be placed on rooftops and utility poles (Cox, 1991, 1995). In a wireless ATM network, where users are globally mobile, the tracking of users is one of the major functions of the wireless network. Each registration area may have a database that is used to support the tracking process (Jain et al., 1999; Marsan, Chiasserini, & Fumagalli, 2001; Rajagopalan, 1995; Siegmund et al., 1995; Simoens et al., 2003).

The ATM-network gateway (large ATM switch) manages the flow of information to and from the wireless ATM network to the wireline ATM networks. The ATM-network gateway is necessary to support connections between the wireline ATM-network users and wireless users, and is responsible for performing location-resolution functionality for wireline network users as described in Jain et al. (1999).

Registration-Area Concept

The wireless ATM network consists of registration areas, the wireless ATM-network backbone, and gateways to the wireline ATM network(s) as depicted in Figure 2. The registration areas of the wireless ATM network are responsible for supporting wireless users.

Each registration area incorporates the signaling functionality required to support mobile users. Via the use of registration areas, the wireless ATM-network architecture is a completely distributed network. By dividing the wireless ATM network into registration areas, the need for addressing the granularity of the wireless ATM network is also reduced. The radio ports and radio-port controllers have only local significance within the registration area. In terms of locating and routing connections to wireless users, the wireless ATM network only considers the registration area of the user and not the particular radio port. In the other direction, the location of the user needs to be updated only when the user moves between the registration areas, which significantly cuts on the signaling traffic.

MOBILITY MANAGEMENT IN WIRELESS ATM NETWORKS

In a wireless ATM network, several procedures are required due to subscriber mobility. Registration is

required to locate a user during information delivery. A connection setup is used to establish connections to other users or servers in the wireless network. Handoff provides true mobility to wireless users and allows them to move beyond the coverage of a single wireless access point. Existing ATM-signaling protocols do not support the registration, connection-setup, and handoff transactions that are required to support wireless users (Lach et al., 2003). In order to support wireless users in the ATM architecture, we need to adapt the registration, connection–setup, and handoff procedures used in existing wireless communication networks (Marsan et al., 2001; Siegmund et al., 1995).

During a study of wireless ATM mobility management, several ideas have been proposed. What is important is to explain the overlay-signaling technique (Chiasserini & Cigno, 2002). Overlay-signaling ATM connections are used to transport mobility-related signaling messages between the registration areas in the wireless ATM network and does not require any changes to the existing ATM protocols. The resulting signaling network is then overlaid on top of the existing ATM network. The motivation for implementing an overlay-signaling network is to remain compatible with the existing ATM protocols. Since there are no modifications to the ATM protocols, the overlay-signaling approach does not require any modifications to the existing ATM infrastructure.

Registration Using Overlay Signaling

Registration is performed to maintain information about the wireless users' locations. It consists of several phases. First, the registration process starts with the transmission of the user identification number and the user's previous registration-area identification from the portable device that enters a new registration area (Cox, 1991; Wang & Lee, 2001). Upon receiving the UID and the authentication information, an ATM connection is established and the user's profile is updated with the new location information. The updated profile is transferred to the current registration area. The user's profile in the previous registration area is deleted by establishing an ATM connection to the previous registration-area switch (PRAS). After the registration transaction is complete, the connection is released.

Session Setup Using Overlay Signaling

The session-setup procedure is used to establish a connection between two wireless network users. The originating registration area refers to the calling user's registration area, and the destination registration area refers to the called user's registration area. The called-user identification number (CUID) is transmitted from the portable device to the originating registration-area switch together with the session-setup parameters such as the required bandwidth, traffic type, and so forth. The originating registration-area switch forms a setup message using the incoming session parameters.

Handoff Using Overlay Signaling

Handoff is the transfer of a user's radio link between radio ports in the network. The portable devices monitor the link quality in terms of received signal power to candidate radio ports, and when the link to another port becomes better, that port is selected and handoff is initiated (Cox, 1991; Wang & Lee, 2001). The link quality is determined by the portable devices because only these devices can determine the quality of the links to multiple radio ports and decide on the best link. In contrast, a radio port can only monitor the link between itself and the portable device. Starting the handoff, the device realizes that a link of better quality exists to a candidate radio port and sends a message to the previous registration-area switch, desiring a handoff to the candidate radio port. The PRAS transfers a copy of the user profile to the candidate registration-area switch (CRAS). the PRAS contacts the end point for the user connection and requests rerouting to the candidate registration area (Cox). Once the rerouting is complete, the PRAS contacts the portable device and relays the channel-assignment information, while the CRAS and the device verify the connection (Cox).

CONCLUSION AND FUTURE TRENDS

In this article, a wireless ATM network is described that can be used in combining future heterogeneous cellular systems (Figure 1). It will expand the range

of offered services and the amount of resources available to wireless users.

The future convergence of several wireless networks in an interoperability environment is critical for the existence and reliability of services worldwide. The interconnection should take special care for mobility procedures, especially for handover, which in our case is considered to be intersystem handover. A common transmission-interconnection network should be implemented, capable of managing all mobility procedures that might take place during the movement and the required services of heterogeneous subscribers. Wireless ATM is a promising candidate since it consists of a robust architecture based on wired ATM, supports multiple services from different sources, and can interconnect different networks as a transport mechanism. The wireless environment poses main problems such as cell losses due to the radio environment, cells out of order in the case of handovers, and general congestion in the case of simultaneous resource demands. Future research on wireless ATM should concentrate on forward error connection (FEC) techniques to guarantee cell integrity, handover algorithms to preserve the cell sequence, call-admission control algorithms to take care of congestion, and priority services and special signaling over existing ATM networks to maintain mobility cases.

REFERENCES

Acampora, A. S. (1994). An architecture and methodology for mobile executed handoff in cellular ATM networks. *IEEE Journal on Selected Areas in Communications, 12*(8), 1365-1375.

ATM Forum. (1996). *ATM Forum user network interface specification version 3.1.*

Chan, K. S., Chan, S., Ko, K. T, Yeung, K. L., & Wong, E. W. M. (2000). An efficient handoff management scheme for mobile wireless ATM networks. *IEEE Transactions on Vehicular Technology, 49*(3), 799-815.

Chiasserini, F. C., & Cigno, R. L. (2002). Handovers in wireless ATM networks: In-band signaling protocols and performance analysis. *IEEE Transactions on Wireless Communications, 1*(1), 87-100.

Cox, D. C. (1991). A radio system proposal for widespread low-power tetherless communications. *IEEE Transactions on Communications, 39*(2), 324-335.

Cox, D. C. (1995). Wireless personal communications: What is it? *IEEE Personal Communications Magazine, 2*(2), 2-35.

Esmailzadeh, R., Nakagawa, M., & Jones, A. (2003). TDD-CDMA for the fourth generation of wireless communications. *IEEE Wireless Communications, 10*(4), 8-15.

Houssos, N., Alonistioti, A., Merakos, L., Mohyeldin, E., Dillinger, M., Fahrmair, M., et al. (2003). Advanced adaptability and profile management framework for the support of flexible mobile service provision. *IEEE Wireless Communications, 10*(4), 52-61.

Jain, R., Rajagopalan, B., & Chang, L. F. (1999). Phone number portability for PCS systems with ATM backbone using distributed dynamic hashing. *IEEE JSAC, 37*(6), 25-28.

Lach, H.-Y., Janneteau, C., & Petrescu, A. (2003). Network mobility in beyond-3G systems. *IEEE Communications Magazine, 41*(7), 52-57.

Marsan, M. A., Chiasserini, C. F., & Fumagalli, A. (2001). Performance models of handover protocols and buffering policies in mobile wireless ATM networks. *IEEE Transactions on Vehicular Technology, 50*(4), 925-941.

Rajagopalan, B. (1995). Mobility management in integrated wireless ATM networks. *Proceedings of Mobicom 1995,* Berkeley, CA.

Siegmund, H., Redl, S. H., Weber, M. K., & Oliphant, M. W. (1995). *An introduction to GSM.* Boston: Artech House.

Simoens, S., Pellati, P., Gosteau, J., Gosse, K., & Ware, C. (2003). The evolution of 5 GHz WLAN toward higher throughputs. *IEEE Wireless Communications, 10*(6), 6-13.

Varquez-Castro, M., Perez-Fontan, F., & Arbesser-Rastburg, B. (2002). Channel modelling for satellite

and HASP system design. *Wireless Communications and Mobile Computing, 2*, 285-300.

Wang, K., & Lee, L. S. (2001). Design and analysis of QoS supported frequent handover schemes in microcellular ATM networks. *IEEE Transactions on Vehicular Technology, 50*(4), 942-953.

KEY TERMS

ATM (Asynchronous Transfer Mode): A transmission technique that transmits combined information in small, fixed-size packets called ATM cells.

B-ISDN (Broadband Integrated-Services Data Network): An ISDN that supports a wider range of voice and nonvoice applications.

EDGE (Enhanced Data for GSM Evolution): An enhanced version of GSM networks for higher data rates. The main difference is the adoption of 8 QPSK (quadrature phase shift keying) modulation in the air interface, which increases the available bit rates.

GPRS (General Packet Radio Services): An evolution of GSM networks that supports data services with higher bit rates than GSM. It uses the same air interface as GSM, but it supports IP signaling back to the core network.

GSM (Global System Mobile): A mobile network that provides all services of fixed telephony to wireless subscribers.

HASP (High-Altitude Stratosphere Platform): A special platform to support overlay coverage in large geographical areas with the advantage of a closer distance than satellites. They operate in the stratosphere at altitudes of up to 22 km, exploiting the best features of both terrestrial and satellite systems. They are usually implemented though the use of unmanned aeronautical vehicles.

MAC (Medium Access Control): A protocol layer above the network layer that provides controlled access to several subscribers that request simultaneous access.

UMTS (Universal Mobile Telecommunication System): The evolution of GSM to higher bandwidth services and multimedia applications.

WLAN (Wireless Local-Area Network): A wireless network that provides access to subscribers with end-to-end IP connections.

HyperReality

Nobuyoshi Terashima
Waseda University, Japan

INTRODUCTION

On the Internet, a cyberspace is created where people communicate together usually by using textual messages. Therefore, they cannot see together in the cyberspace. Whenever they communicate, it is desirable for them to see together as if they were gathered at the same place. To achieve this, various kinds of concepts have been proposed such as a collaborative environment, tele-immersion, and telepresence (Sherman & Craig, 2003).

In this article, HyperReality (HR) is introduced. HR is a communication paradigm between the real and the virtual (Terashima, 2002; Terashima, 1995; & Terashima & Tiffin, 2002). The real means a real inhabitant such as a real human or a real animal. The virtual means a virtual inhabitant, a virtual human or a virtual animal.

HR provides a communication environment where inhabitants, real or virtual, those are at different locations, come, see, and do cooperative work together as if they were gathered at the same place. HR can be developed based on Virtual Reality (VR)and telecommunications technologies.

BACKGROUND

VR is a medium composed of interactive computer simulations that sense the viewer's position and actions and replace or augment the feedback to one or more senses such as seeing, hearing, and/or touch, giving the feeling of being mentally immersed or present in the virtual space (Sherman and Craig, 2003). They can have a stereoscopic view of the object and its front view or side view according to their perspectives. They can touch and/or handle the virtual object by hand gesture (Burdea, 2003; Kelso, 2002; Stuart 2001).

Initially, computer-generated virtual realities were experienced by individuals at single sites. Then, sites were linked together so that several people could interact in the same virtual reality. The development of the Internet and broadband communications now allows people in different locations to come together in a computer-generated virtual space and to interact to carry out cooperative work.

This is collaborative virtual environment. As one of collaborative environments, the NICE project has been proposed and developed. In this system, children use avatars to collaborate in the NICE VR application, despite being at geographically different locations and using different styles of VR systems (Johnson, Roussos, Leigh, Vasilakis, Marnes & Moher, 1998). A combat simulation and VR game are applications of collaborative environment.

Tele-Immersion (National Tele-Immersion Initiative=NTII) will enable users at geographically distributed locations to collaborate in real time in a shared, simulated environment as if they were in the same physical room (Lanier,1998).

HR provides a communication means between real inhabitants and virtual inhabitants, as well as a communication means between human intelligence and artificial intelligence. In HR, communication paradigm for the real and the virtual is defined clearly. Namely, in HR, a HyperWorld (HW) and coaction fields (CFs) are introduced.

Augmented Reality (AR) is fundamentally about augmenting human perception by making it possible to sense information not normally detected by the human sensory system (Barfield & Caudell, 2001). A 3D virtual reality derived from cameras reading infrared or ultrasound images would be AR. A 3D image of a real person based on conventional camera imaging that also shows images of their liver or kidneys derived from an ultrasound scan is also a form of AR. HR can be seen as including AR in the sense that it can show the real world in ways that humans do not normally see it. In addition to this, HR provides a communication environment between the real and the virtual.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

HR CONCEPT

The concept of HR, like the concepts of nanotechnology, cloning and artificial intelligence, is in principle very simple. It is nothing more than the technological capability to intermix VRwith physical reality (PR) in a way that appears seamless and allows interaction. HR incorporates collaborative environment(Sherman,2003), but it also links collaborative environment with the real world in a way that seeks to be as seamless as possible. In HR, it is the real and virtual elements which interact and in doing so they change their position relative to each other. Moreover, the interaction of the real and virtual elements can involve intelligent behavior between the two and this can include the interaction of human and artificial intelligence. However, HR can be seen as including AR in the sense that it can show the real world in ways that humans do not normally see it.

HR is made possible by the fact that, using computers and telecommunications, 2D images from one place can be reproduced in 3D virtual reality at another place. The 3D images can then be part of a physically real setting in such a way that physically real things can interact synchronously with virtually real things. It allows people not present at an actual activity to observe and engage in the activity as though they were actually present. The technology will offer the experience of being in a place without having to physically go there. Real and virtual objects will be placed in the same "space" to create an environment called an HW. Here, virtual, real, and artificial inhabitants and virtual, real, and artificial objects and settings can come together from different locations via communication networks, in a common place of activity called a CF, where real and virtual inhabitants can work and interact together.

Communication in a CF will be by words and gestures and, sometimes, by touch and body actions. What holds a CF together is the domain knowledge which is available to participants to carry out a common task in the field. The construction of infrastructure systems based on this new concept means that people will find themselves living in a new kind of environment and experiencing the world in a new way.

HR is still hypothetical. Its existence in the full sense of the term is in the future. Today parts of it have a half-life in laboratories around the world. Experiments which demonstrate its technical feasibility depend upon high-end work stations and assume broad-band telecommunications. These are not yet everyday technologies. HR is based on the assumption that Moore's law will continue to operate, that computers will get faster and more powerful and communication networks will provide megabandwidth.

The project that led to the concept of HR began with the idea of virtual space teleconferencing system. It was one of the themes of ATR (Advanced Telecommunications Research) in Kansai Science City. Likened to the Media Lab at MIT or the Santa Fe Institute, ATR has acquired international recognition as Japan's premier research centre concerned with the telecommunication and computer underpinnings of an information society. The research lasted from 1986 to 1996 and successfully demonstrated that it was possible to sit down at a table and engage interactively with the telepresences of people who were not physically present. Their avatars looked like tailor's dummies and moved jerkily. However, it was possible to recognise who they were and what they were doing and it was possible for real and virtual people to work together on tasks constructing a virtual Japanese portable shrine by manipulating its components (Terashima, 1994).

The technology involved comprised two large screens, two cameras, data gloves, and glasses. Virtual versions were made of the people, objects, and settings involved and these were downloaded to computers at different sites before the experiment's start. Then it was only necessary to transmit movement information of positions and shapes of objects in addition to sound. As long as one was orientated toward the screen and close enough not to be aware of its edges, inter-relating with the avatars appeared seamless. Wearing a data glove, a viewer can handle a virtual object by hand gesture. Wearing special glasses, he/she can have a stereoscopic view of the object.

Most humans understand their surroundings primarily through their senses of sight, sound, and touch. Smell and even taste are sometimes critical too. As well as the visual components of physical and virtual reality, HR needs to include associated sound, touch, smell, and taste. The technical challenge of HR is to make physical and virtual reality appear to the full human sensory apparatus to intermix seamlessly. It is not dissimilar to, or disassociated from, the challenges

that face nanotechnology at the molecular level, cloning at the human level and artificial intelligence at the level of human intelligence. Advanced forms of HR will be dependent on extreme miniaturisation of computers. HR involves cloning, except that the clones are made of bits of information. Finally, and as one of the most important aspects of HR, it provides a place for human and artificial intelligences to interact seamlessly. The virtual people and objects in HR are computer-generated and can be made intelligent by human operation or they can be activated by artificial intelligence.

HR makes it possible for the physically real inhabitants of one place to purposively coact with the inhabitants of remote locations as well as with other computer-generated artificial inhabitants or computer agents in an HW.

An HW is an advanced form of reality where real-world images are systematically integrated with 3D images derived from reality or created by computer graphics. The field of interaction of the real and virtual inhabitants of an HW is defined as a CF.

An example of HR is shown in Figure 1. In Figure 1, a virtual girl is showing her virtual balloon to a real girl in CFa. Two adults, one real and one virtual are discussing something in CFb which is a coaction field for interpreting between Japanese and English. They must be able to speak either Japanese or English. A real boy is playing ball with a virtual puppy in CFc. The boy and the puppy share the knowledge of how to play ball.

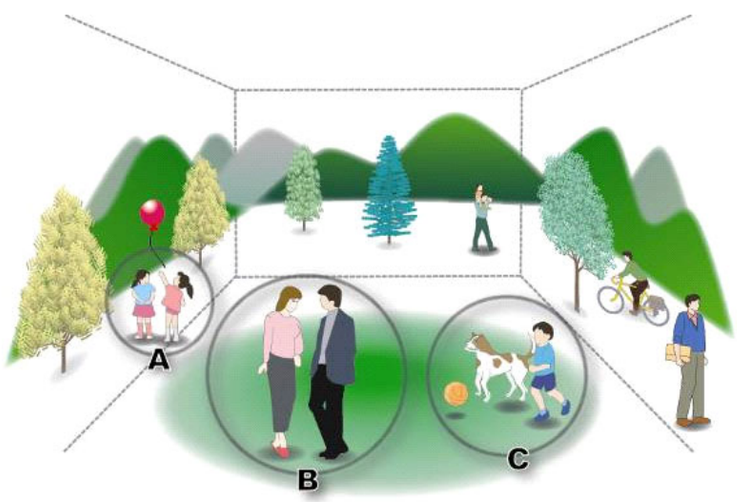
HyperWorld

An HW is a seamless integration of a (physically) real world (RW) and a virtual world (VW). HW can, therefore, be defined as (RW, VW).

A real world consists of real natural features such as real buildings and real artifacts. It is whatever is atomically present in a setting and is described as (SE), that is, the scene exists. A virtual world consists of the following:

- **SCA (scene shot by camera):** Natural features such as buildings and artifacts that can be shot with cameras (video and/or still), transmitted by telecommunications and displayed in VR.
- **SCV (scene recognised by computer vision):** Natural features such as buildings ,arti-

Figure 1. An example of HyperReality



facts, and inhabitants whose 3D images are already in a database are recognized by computer vision, transmitted by telecommunications and reproduced by computer graphics and displayed in VR.

- **SCG (scene generated by computer graphics):** 3D Objects created by computer graphics, transmitted by telecommunications and displayed in VR. SCA and SCV refer to VR derived from referents in the real world whereas SCG refers to VR that is imaginary. A VW is, therefore, described as: (SCA, SCG, SCV). This is to focus on the visual aspect of a HW. In parallel, as in the real world, there are virtual auditory, haptic, and olfactory stimuli derived either from real world referents or generated by computer.

Coaction Field

A CF is defined in an HW. It provides a common site for objects and inhabitants derived from PR and VR and serves as a workplace or an activity area within which they interact. The CF provides the means of communication for its inhabitants to interact in such joint activities as designing cars or buildings or playing games. The means of communication include words, gestures, body orientation, and movement, and in due course will include touch. Sounds that provide feedback in performing tasks, such as a reassuring click as elements of a puzzle lock into place or as a bat hits a ball, will also be included. The behaviour of objects in a CF conforms to physical laws, biological laws or to laws invented by humans. For a particular kind of activity to take place between the real and virtual inhabitants of a CF, it is assumed that there is a domain of knowledge based on the purpose of the CF and that it is shared by the inhabitants.

Independent CFs can be merged to form a new CF, termed the outer CF. For this to happen, an exchange of domain knowledge must occur between the original CFs, termed the inner CFs. The inner CFs can revert to their original forms after interacting in an outer CF. So, for example, a CF for designers designing a car could merge with a CF for clients talking about a car which they would like to buy to form an outer CF that allowed designers to exchange information about car with clients. The CF for exchanging information between designers and clients would terminate and the outer CF would revert to the designers' CF and clients' CF.

A CF can therefore be defined as:

CF={field, inhabitants (n>1), means of communication, knowledge domain, laws, controls}

In this definition, a field is the locus of the interaction which is the purpose of the CF. This may be well defined and fixed as in the baseball field of a CF for playing baseball or the golf course of a CF for playing golf. Alternatively, it may be defined by the action as in a CF for two people walking and talking, where it would be opened by a greeting protocol and closed by a goodbye protocol and, without any marked boundary, would simply include the two people. Inhabitants of a CF are either real inhabitants or virtual inhabitants. A real inhabitant (RI) is a real human, animal, insect, or plant. A virtual inhabitant (VI) consists of the following:

- **ICA (inhabitant shot by camera):** Real people, animals, insects or plants shot with cameras, (transmitted) and displayed in VR.
- **ICV (inhabitant recognised by computer vision):** Real people, animals, insects, or plants recognised by computer vision, (transmitted), reproduced by computer graphics and displayed using VR.
- **ICG (inhabitant generated by computer graphics):** An imaginary or generic life form created by computer graphics, (transmitted) and displayed in VR.

VI is described as: (ICA, ICG, ICV).

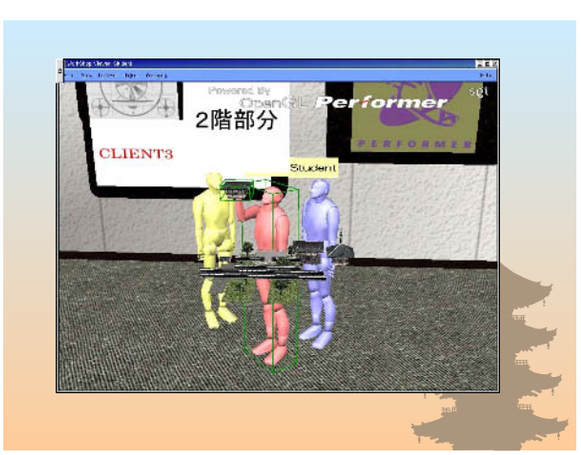
Again we can see that ICA and ICV are derived from referents in the real world, whereas an ICG is imaginary or generic. By generic, we mean some standardised, abstracted non-specific version of a concept, such as a man, a woman, or a tree. It is possible to modify VR derived from RW or mix it with VR derived from SCG. For example, it would be possible to take a person's avatar which has been derived from their real appearance and make it slimmer, better-looking, and with hair that changes colour according to mood. Making an avatar that is a good likeness can take time. A quick way is to take a standard body and, as it were, paste on it a picture of a person's face derived from a photo.

An ICG is an agent that is capable of acting intelligently and of communicating and solving problems. The intelligence can be that of a human referent or it can be an artificial intelligence based on automatic learning, knowledge base, language understanding, computer vision, genetic algorithm, agent, and image processing technologies. The implications are that a CF is where human and artificial inhabitants communicate and interact in pursuit of a joint task.

The means of communication relates to the way that CFs in the first place would have reflected light from the real world and projected light from the virtual world. This would permit communication by written words, gestures, and such visual codes as body orientation and actions. It would also have sound derived directly from the real world and from a speaker linked to a computer source which would allow communication by speech, music, and coded sounds. Sometimes it will be possible to include haptic and olfactory codes.

The knowledge domain relates to the fact that a CF is a purposive system. Its elements function in concert to achieve goals. To do this there must be a shared domain of knowledge. In a CF this resides within the computer-based system as well as within the participating inhabitants. A conventional game of tennis is a system whose boundaries are defined by the tennis court. The other elements of the system, such as balls and rackets, become purposively interactive only when there are players who know the object of the game and how to play it. Intelligence resides in the players. However, in a virtual game of tennis all the elements, including the court, the balls the racquets and the net, reside in a database. So too do the rules of tennis. A CF for HyperTennis combines the two. The players must know the game of tennis and so too must the computer-based version of the system. This brings us to the laws in a CF. These follow the laws of humans and the laws of nature. By the laws of nature are meant the laws of physics, biology, electronics, and chemistry. These are of course given in that part of a CF which pertains to the real world. They can also be applied to the intersecting virtual world, but this does not necessarily have to be the case. For example, moving objects may behave as they would in physical reality and change shape when they collide. Plants can grow, bloom, seed, and react to sunlight naturally. On the other hand, things can fall upwards in VR and plants can be programmed to grow in response to music. These latter are examples of laws devised by humans which could be applied to the virtual aspect of a CF.

Figure 2. Scene of HyperClass



HR APPLICATIONS

The applications of HR would seem to involve almost every aspect of human life, justifying the idea of HR becoming an infrastructure technology. They range from providing home care and medical treatment for the elderly in ageing societies, to automobile design, global education and HyperClass (Rajasingham, 2002; Terashima & Tiffin, 1999, 2000; Terashima, Tiffin, Rajasingham, & Gooley, 2003; Tiffin, 2002; Tiffin & Rajasingham, 2003; Tiffin, Terashima, & Rajasingham, 2001), city planning (Terashima, Tiffin, Rajasingham, & Gooley, 2004),games and recreational activities and HyperTranslation (O'Hagan 2002).

A scene of HyperClass is shown in Figure 2. In this figure, three avatars are shown: one (center) is a teacher. It handles a part of Japanese virtual shrine. The other two are students. They are watching the operation.

FUTURE FORECAST

HR waits in the wings. For HR to become the information infrastructure of the information society, we need a new generation of wearable personal computers with the processing power of today's mainframe and universal telecommunications where bandwidth is no longer a concern. Such conditions should obtain sometime in 10 to 20 years. Now, a PC-based HR platform and screen-based HR are available.

In 10 years, a room based HR will be developed. In 20 years, universal HR will be accomplished. In this stage, they will wear intelligent data suits which provide a communication environment where they come, see, talk, and cooperate together as if they were at the same place.

CONCLUSION

Virtual reality is in its infancy. It is comparable to the state of radio transmission in the last year of the 19th century. It worked, but what exactly was it and how could it be used? The British saw radio as a means of contacting their navy by Morse code and so of holding their empire together. No one in 1899 foresaw its use first for the transmission of voices and music and then for television. Soon radio will be used for transmitting virtual reality and one of the modes of HR in the future will be based on broadband radio transmissions.

This chapter has tried to say what HR is in terms of how it functions and how it relates to other branches of VR. HR is still in the hands of the technicians and it is still in the laboratory for improvement after trials. But a new phase has just begun. HR is a medium and the artists have been invited in to see what they can make of it.

REFERENCES

Barfield, W. & Caudell, T. (2001). *Fundamentals of wearable computers and augmented reality*. Lawrence Erlbaum.

Burdea, G. & Coiffet, P. (2003). *Virtual reality technology* (2nd Ed.). John Wiley & Sons.

Johnson, A., Roussos, M., Leigh, J., Vasilakis, C., Marnes, C. & Moher, T. (1998).The Nice Project: Learning together in a virtual world. *Proceedings of the IEEE1998 Virtual Reality Annual Conference* (pp. 176-183).

Kelso, J., Lance, A., Steven, S. & Kriz, R. (2002). DIVERSE: A framework for building extensible and reconfigurable device-independent virtual environments. *Proceedings of the IEEEVirtual Reality 02.*

Lanier, J. (1998). National tele-Immersion Initiative. Online *http://www.advanced.org/teleimmersion.html*

O'Hagan, M. (2002). *HyperTranslation: HyperReality-paradigm for the third millenium*. UK: Routledge.

Rajasingham, L. (2002). Virtual class and HyperClass: Interweaving pedagogical needs and technological possibilities. *Groningen Colloquium on Language Use and Communication,* CLCG.

Sherman, W. & Craig, A. (2003). *Understanding virtual reality interface, application and design.*

Stuart, R. (2001). *Design of virtual environment*. Barricade Books.

Terashima, N. (2002). *Intelligent communication systems.* Academic Press.

Terashima, N. (1995). HyperReality. *Proceeding of the International Conference on Recent Advance in Mechatronics* (pp. 621-625).

Terashima, N. (1994). Virtual space teleconferencing system-distributed virtual environment. *Proceedings of the 3rd International Conference on Broadband Islands* (pp. 35-45).

Terashima, N. & Tiffin, J. (2002). *HyperReality: Paradigm for the third millennium.* Routledge.

Terashima, N. & Tiffin, J. (2000). HyperClass. *Open Learning 2000 Conference Abstracts.*

Terashima, N. & Tiffin, J. (1999). An experiment of virtual space distance learning systems. *Proceedings Annual Conference of Pacific Telecommunication Council* (CD-Rom).

Terashima, N., Tiffin, J., Rajasingham, L. & Gooley, A. (2003). HyperClass: Concept and its experiment. *Proceedings of PTC2003* (CD-Rom).

Terashima, N., Tiffin, J., Rajasingham, L. & Gooley, A. (2004). Remote collaboration for city planning. *Proceedings of PTC2004* (CD-Rom).

Tiffin, J. (2002). The HyperClass: Education in a broadband Internet environment. *Proceedings of the International Conference on Computers in Education* (pp. 23-29).

Tiffin, J. & Rajasingham, L. (2003). *Global virtual university.* Routledge Farmer.

Tiffin, J., Terashima, N. & Rajasingham, L. (2001). Artificial Intelligence in the HyperClass: Design issues. *Computers and Education Towards an Interconnected Society*, 1-9.

KEY TERMS

Augmented Reality: Intermixing a physical reality and a virtual reality.

Coaction Field: A place where inhabitants, real or virtual, work or play together as if they were gathered at the same place.

HyperClass: Intermixing a real classroom and a virtual classroom where a real teacher and students and a virtual teacher and students come together and hold a class.

HyperReality: Providing a communication environment where inhabitants, real or virtual, at different locations, are brought together through the communication networks and work or play together as if they were at the same place.

HyperWorld: Intermixing a real world and a virtual world seamlessly.

Remote Collaboration: They come together as their avatars through the communication networks as if they were gathered at the same place.

Virtual Reality: Simulation of a real environment where they can have feelings of seeing, touch, hearing, and smell.

Improving Student Interaction with Internet and Peer Review

Dilvan de Abreu Moreira
University of São Paulo, Brazil

Elaine Quintino da Silva
University of São Paulo, Brazil

INTRODUCTION

In the last few years, education has gone through an important change—the introduction of information technology in the educational process. Many efforts have been conducted to realize the benefits of technologies like the Internet in education. As a result of these efforts, there are many tools available today to produce multimedia educational material for the Web, such as WebCT (WebCT, 2004). However, teachers are not sure how to use these tools to create effective models for teaching over the Internet. After a teacher puts classroom slides, schedules, and other static information in his or her Web pages, what more can this technology offer? A possible response to this question is to use Internet technologies to promote collaborative learning.

Collaborative Learning (CL) is an educational strategy based on social theories in which students joined in small groups are responsible for the learning experience of each other (Gokhale, 1995; Panitz, 2002). In CL, the main goal of the teacher is to organize collective activities that can stimulate the development of skills such as creativity, oral expression, critical thinking, and others. When supported by computers and Internet technologies, collaborative learning is referenced as Computer Supported Collaborative Learning (CSCL). The main goal of CSCL is to use software and hardware to support and increase group work and learning. The peer review method, known by almost everyone in the academic world, when applied as an educational tool, can be considered a kind of collaborative learning activity.

This article describes an educational method that uses peer review and the Internet to promote interaction among students. This method, which has been used and refined since 1997 (by the first author), is used currently in different computer science courses at the ICMC-USP. A software tool—the WebCoM, Web Course Manager (Silva & Moreira, 2003)—is also presented. It supports the peer review method to improve interaction among students. The main advantages of the use of the peer review method and the WebCoM tool over other works in this context are that they:

- allow debate between groups (workers and reviewers) to improve interaction and social abilities among students;
- focus on the interaction among students and their social skills; and
- offer support for group activities (such as reports and assignments) without peer review.

Results generated by the experience of managing classes with the WebCoM tool are also presented.

THE STUDENT GROUPS WITH PEER REVIEW METHOD

The peer review process is commonly used in the academic world; an article, project, or course is proposed, and peers judge the merits of the work. It is used in the educational context with a variety of goals, but almost always it is focused on communication and writing skills (Helfers et al., 1999; Kern et al., 2003; Nelson, 2000).

In the educational peer review method presented here, students join in groups to carry out an assignment. After that, each assignment is made public using the Internet and is judged by another group of fellow students. These reviewers write a review report presenting their opinions about the work. Once the

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

reviewers' work becomes public, the teacher schedules a class debate. At this debate, each group presents its work and has a chance to defend it from the criticisms of the reviewers. The two groups debate the work in front of their classmates and teacher. Usually, the teacher is able to grade the assignment based on the review and the debate.

Trying to do all these tasks by hand would greatly reduce the benefits of the method because of the work needed to implement it. A software tool is necessary to manage the assignment process. So, few authors have developed Web-based software to assist it, such as:

- The PG (Peer Grader), a system that offers support to peer review activities in which students submit work, review other works, and grade the reviewed work. The final grade of each work is determined by the system, based on the grade of the reviewers (Gehringer, 2000).
- The WPR (Web-Based Peer Review) system is mentioned in Liu et al. (2001) as a tool for peer review management. Although some results of experiments using this tool are presented, there is only a brief explanation about the tool and no references to specific information about the system.

There are other Web-based tools that can be adapted for classroom use, such as CyberChair (CyberChair, 2004) and WIMPE (Nicol,1996), which support the review process for technical contributions to conferences.

The major problem with these tools is the kind of review they support, one not targeted to promote interaction in educational environments. A new tool, WebCoM (Web Course Manager), was developed specifically to address this issue. Its main objective is to provide graphical interfaces to get, store, manipulate, and present information generated by both student groups and teachers during a course. Using the WebCoM tool, the teacher can:

- define assignments and deadline dates;
- define other activities such as reports and tests;
- define which group a reviewer will review; and
- associate grades to students or groups.

The students can:

- create groups;
- turn in assignments and reports;
- view and access works of others groups; and
- access their grades.

As a practical example, the next section shows how a very common kind of assignment for computer science courses—a software project—can be handled using WebCoM and peer review to promote interaction among students.

SOFTWARE PROJECT ASSIGNMENT

The software project is a classic assignment in computer science courses. Commonly, in this type of activity, students are required to put into practice all concepts taught in class. There are two ways to conduct the software project activity: first, all students (or groups) develop a project from the same subject; and second, each student (or group) develops a project from a different subject. In either way, students are limited to explore and learn only about the project they are working on, mainly because of the individualism from traditional education methods (Panitz & Panitz, 1998). The presented peer review method minimizes this limitation, because students (or groups) are required to learn about their colleagues' projects. When required to review projects and to participate in debates about other projects, students have an opportunity to extend their knowledge about other subjects, expanding the experience they would have using traditional individual learning.

The development of a software project under the peer review process has five steps: group formation, assignment upload, choosing review groups, review upload, and classroom debate.

At the beginning of the course, the students have access to the course Web pages, where they can find the usual material (lecture slides, course calendar, etc.) and a list of available software projects. These projects are previously defined by the teacher and relate to the subject being taught in the course. In addition, students have access to the WebCoM tool, in which the course and its activities (assignments

and projects) are registered. The next subsections describe each of the five steps of the process.

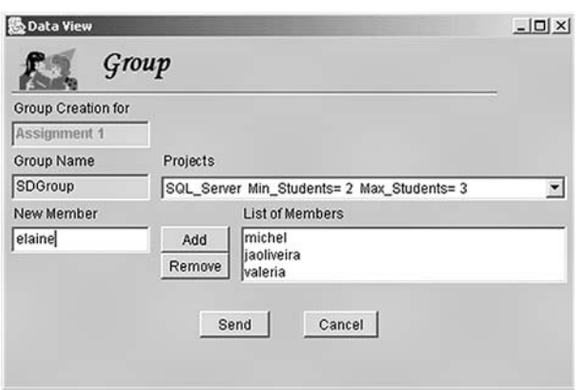
Group Formation

After signing into the WebCoM tool, students have to form groups, usually three to four components. At this stage they can choose which project they want to work on. There are a limited number of projects, and each one can be worked on by a limited number of groups. As the groups are formed, the options are reduced in a first-come, first-served basis. After the group creation, the management tool creates an area on the server to store files uploaded by the groups (assignment and review report). Figure 1 shows the interface of the WebCoM's Group Formation tool.

Assignment Upload

Until the deadline, groups can upload their work as many times as they wish, using the WebCoM FTP tool. It automatically defines where to put the uploaded files, based on the group from which the logged student is a member. The use of a software tool is important at this point, because once the files are uploaded, they can be organized in Web pages and accessed by reviewers. Soon after the upload, the files are made available on a WebCoM HTML page (Figure 2).

Figure 1. WebCoM group formation tool (reproduced with permission from E.Q. Silva and D.A. Moreira, ACM JERIC 3:1-14, Nov. 2003. Association for Computing Machinery)



Specifically for the software project, students have to upload the code and a structured report called UDF (Unit Development Folder) (Williams, 1975). Other kinds of structured reports can be used, but it is important to have a structured report about the code being uploaded. That report is used to normalize the review process.

Choosing Review Groups

After the deadline for hand-in (upload) of the assignments, the teacher can determine which group another group will review. The teacher can take this opportunity to pair complementary projects, avoid cross reviews (two groups doing the review of each other), or any other strategy the teacher thinks may improve the quality of the reviews and the final debate. This task also can be done using a WebCoM tool for review allocation.

Review Upload

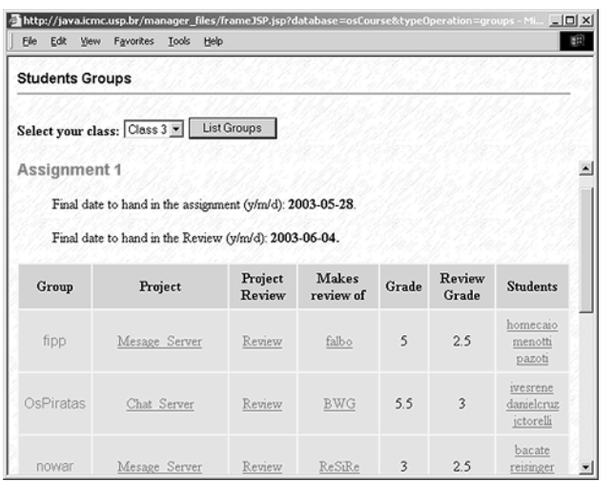
Until the deadline for the review, the reviewer groups can upload their work as many times as they wish, using the WebCoM FTP tool. Again the tool automatically sets the directory to upload files, based on the logged student information, and makes files available on a WebCoM HTML page (Figure 2).

Reviewers have to test the programs and read the reports about their colleague projects. At this stage, reviewers are encouraged to iterate with the group that did the work in order to better understand the project and exchange ideas. After that, they try to answer specific questions in their review; for instance, design quality, code quality, and documentation quality. It is important that judging parameters for each question are clearly defined to the students.

Classroom Debate

That is the most interesting part of the method. In the classroom (or in a chat room for distance education courses), each group has a chance to present its work to its classmates (and teacher) and to defend it against the reviewers' criticisms. The correspondent reviewer group can present its suggestions and defend its points of view. The two groups can debate the project problems and qualities for some time. Teacher

Figure 2. WebCoM tool for viewing assignments results (reproduced with permission from E.Q. Silva)and D.A. Moreira, ACM JERIC 3:1-14, Nov. 2003. Association for Computing Machinery)


http://java.icmc.usp.br/manager_files/frameJSP.jsp?database=osCourse&typeOperation=groups - Mi...
File Edit View Favorites Tools Help
Students Groups
Select your class: Class 3 List Groups
Assignment 1
Final date to hand in the assignment (y/m/d): 2003-05-28.
Final date to hand in the Review (y/m/d): 2003-06-04.
Group Project Project Review Makes review of Grade Review Grade Students
fipp Mesage_Server Review falbo 5 2.5 homecaio menotti pazoti
OsPiratas Chat_Server Review BWG 5.5 3 wesrene danielcruz ictorelli
nowar Mesage_Server Review ReSiRe 3 2.5 bacate reisinger

and classmates can give opinions, ask questions, and contribute to the debate. The process goes on until all groups have presented their work. Usually, the teacher can give a grade to the groups, based on the reviews and the debates. During the debates, it is easier to notice if a group really understood the theory and key concepts behind its software project.

It is recommended that the teacher plan the course schedule to leave sufficient time for the debates. Some groups debate more than others. If the time for debate is too short, the students will not have time to expose their points of view.

At the end of the process, all information is made available in an organized way at the course site. Figure 2 shows a WebCoM page that summarizes the results of an activity managed with the peer review method.

In Figure 2, *Group* is the name of the group; *Project* is a link to the assignment done by the group; *Project Review* is a link to the review of the group's project; *Makes review of* is a link to the review written by the group; *Grade* is the grade for the project; *Review Grade* is the grade for the review; and *Students* are the members of the group.

The example of a software project assignment describes well how the method works, but this method has been used in other kinds of assignments. When used in seminar assignments, where groups have to present a seminar about a subject to the class, the review strategy is slightly modified. The groups upload the text and slides they intend to present, and then the reviewers (usually after a week) upload their opinions. Now the groups have the chance to modify their text and slides, based upon the opinions of the reviewers, if they agree with them. After the seminar presentation, there is the debate between the group and the reviewers (the audience is invited to take part, too) where the reviewers can present their opinions about the seminar presentation, analyze if the modifications they proposed were properly implemented (if they were accepted), and point out the qualities and problems of the work. Again, the group is free to challenge the opinions of the reviewers. This strategy improves the quality of the seminars and helps to start a good debate about the seminar.

TESTING THE PEER REVIEW METHOD IN THE REAL WORLD

This method of student groups and peer review has been in use and refinement since 1997, with good results. Since August 2001, the method has been evaluated using the student evaluation questionnaire for graduate and undergraduate courses. To get a picture of how the participating students were seeing the peer review method and WebCoM tool, the following questions, from the student evaluation questionnaire were analyzed:

- **Question 1:** Did you use the WWW facilities? (Y/N)
- **Question 3:** Does the use of the WWW facilities make the course easier? (Y/N)
- **Question 7:** What is your opinion about the idea of Internet support?
- **Question 9:** What do you think of peer review evaluation?

The questionnaire was applied to seven classes from graduate and undergraduate courses, two from the second semester of 2001, two from the second semester of 2002, two from the first semester of 2003, and one from the first semester of 2003. Table

Table 1. Answered questionnaires

	Graduate Students		Undergraduate Students	
	Total	Answered	Total	Answered
2nd Semester 2001	32	18 or ~56%	40	30 or ~75%
2nd Semester 2002	24	22 or ~92%	48	31 or ~65%
1st Semester 2003	24	20 or ~83%	42	34 or ~81%
2nd Semester 2003	-	-	37	29 ou ~78%
Total	80	60 or ~75%	167	124 or ~74%

1 shows the total number of students in each class and the total number of students that answered the questionnaire.

Three persons—a teacher, a psychologist, and a graduate student—classified the student answers in three categories: *Yes or Liked*, *Neutral,* and *No or Disliked*, based upon what the students were asked. The three classifications were merged into one, using averages. Question 3 was used just to make sure all students used the WebCoM tool. Table 2 shows the results of this evaluation (the percentages were calculated taking only the students that answered the questions).

As shown on Table 2, few students disliked the use of the Internet in general (Questions 3 and 7). The majority of the students (both graduate and undergraduate) had a good response to the peer review method (Question 8). Also interesting are the topics raised by the students in their answers about the peer review method/WebCoM (Question 8):

- **Interaction:** 13% graduate and 21% undergraduate students stated in their answers that the method increased interaction or that they learned more about the project of the group they reviewed.
- **Fairness:** 21% graduate and 5% undergraduate students were concerned about having clear judging parameters. As the students are doing the evaluation, they are concerned that different reviewers may be using different parameters for their evaluation. This highlights the need for clear judging parameters being explained in advance by the teacher. Thus, if a group thinks its reviewers did not stick to these parameters, they can bring up the issue during the debate.
- **Embarrassment:** 26% graduate and 6% undergraduate students felt that the review process caused friction among students or that they were embarrassed or uneasy during the debates. They were not comfortable exposing their work and/or receiving criticisms. However, these students are having an opportunity to learn how to overcome those feelings. This is important, as they will be exposed to criticism from their peers throughout their careers.

CONCLUSION

This method of student groups with peer review is one way to explore the real potential of the Internet as an educational tool. The method uses the communication capabilities of the Internet to stimulate more interaction among the students, create an environment to foster constructive debate (collaborative learning), give the students a chance to learn how to give and receive criticism in a polite and constructive way, and provide an engaging environment for the participants (very helpful with dull topics).

Table 2. Answers to the four questions in both years

	Graduate			Undergraduate		
	Yes or liked.	Neutral	No or disliked	Yes or liked.	Neutral	No or disliked
Question 3	90%	3%	7%	82%	9%	9%
Question 7	90%	6%	4%	93%	6%	1%
Question 8	71%	9%	21%	81%	8%	10%

The role of a software tool such as WebCoM in managing the peer review method activities is key to the success of the process as a whole. The method can help the students learn how to:

- present their work, because they have to show their results and opinions to another group and to the rest of the class; therefore, they have to learn how to convince people about a subject;
- evaluate the quality of the work of others, because they have to present constructive criticisms about them; and
- accept and understand criticisms from their colleagues, which is very important for a successful computer science professional.

Teachers can save time by letting part of the evaluation work be done by students. This extra time can be used to manage more groups of students (with less students per group) or to focus on problematic students, who may need extra help.

The main negative point of this method is that some students let personal involvement interfere when they receive criticisms from fellow students. However, this is something that students should begin to change when they are still at school rather than when they become computer science professionals.

REFERENCES

Gehringer, E.F. (2000). Strategies and mechanisms for electronic peer review. *Proceedings of the Frontiers in Education Conference, 30th Annual,* Kansas City, MO.

Gokhale, A.A. (1995). Collaborative learning enhances critical thinking. *Journal of Technology Education*, *7*(1). Retrieved July 20, 2004, from *http://scholar.lib.vt.edu/ejournals/JTE/*

Helfers, C., Duerden, S., Garland, J., & Evans, D.L. (1999). An effective peer revision method for engineering students in first-year English courses. *Proceedings of the Frontiers in Education Conference, 29th Annual,* San Juan, Puerto Rico.

Kern, V.M., Saraiva, L.M., & Pacheco, R.C.S. (2003). Peer review in education: Promoting collaboration, written expression, critical thinking, and professional responsibility. *Education and Information Technologies—Journal of the IFIP Technical Committee on Education*, *8*(1), 37-46.

Liu, E.Z., Lin, S.S.J., Chiu, C., & Yuan, S. (2001). Web-based peer review: The learner as both adapter and reviewer. *IEEE Trans. Education, 44*(3), 246-251.

Nelson, S. (2000). Teaching collaborative writing and peer review techniques to engineering and technology undergraduates. *Proceedings of the Frontiers in Education Conference, 30th Annual*, Kansas City, MO.

Nicol, D.M. (1996). Conference program management using the Internet. *IEEE Computer, 29*(3), 112-113.

Panitz, T. (2002). Using cooperative learning to create a student-centered learning environment. *The Successful Professor*, *1*(1). Millersville, MD: Simek Publishing. Online *www.thesuccessfulprofessor.com*

Panitz, T., & Panitz, P.(1998). Encouraging the use of collaborative teaching in higher education. In J. Forest (Ed.), *University teaching: International perspectives* (pp. 161-202). New York: RoutledgeFalmer Press.

Silva, E.Q., & Moreira, D.A. (2003). WebCoM: A tool to use peer review to improve student interaction. *ACM Journal on Education Resources in Computing, 3*(1), 1-14.

van de Stadt, R. (2004). CyberChair: A Web-based paper submission and reviewing system. Retrieved June 10, 2004, from *http://www.cyberchair.org*

WEBCT Software. (n.d.). Retrieved August 25, 2004, from *http://www.webct.com*

Williams, R.D. (1975). Managing the development of reliable software. *Proceedings of the International Conference on Reliable Software* (pp. 3-8), Los Angeles, California.

KEY TERMS

Collaborative Learning: An instruction method in which students work in groups toward a common academic goal.

CSCL: Computer Supported Collaborative Learning is a research area that uses software and hardware to provide an environment for collaborative learning**.**

FTP: File Transfer Protocol is a protocol to transfer files from one computer to another over the Internet.

Individual Learning: An instruction method in which students work individually at their own level and rate toward an academic goal.

Peer Review Method: Peer review is a process used for checking the work performed by one's equals (peers) to ensure it meets specific criteria. The peer review method uses peer review to evaluate assignments from student groups.

Software Project: An educational activity in which students are required to develop or specify a program following guidelines and requirements that were previously established.

UDF: Unit Development Folder is a kind of structured report to describe a development process.

Information Hiding, Digital Watermarking and Steganography

Kuanchin Chen
Western Michigan University, USA

INTRODUCTION

Digital representation of data is becoming popular as technology improves our ways of information dissemination, sharing and presentation. Without careful planning, digitized resources could easily be misused, especially those distributed broadly over the Internet. Examples of such misuse include use without owner's permission and modification of a digitized resource to fake ownership. One way to prevent such behaviors is to employ some form of authentication mechanism, such as digital watermarks.

Digital watermarks refer to data embedded into a digital source (e.g., images, text, audio or video recording). They are similar to watermarks in printed materials, as a message inserted in the source typically becomes an integral part of the source. Apart from traditional watermarks in printed forms, digital watermarks may be: invisible, in forms other than graphics and digitally removed.

INFORMATION HIDING, STEGANOGRAPHY AND WATERMARKING

To many people, information hiding, steganography and watermarking refer to the same set of techniques to hide some form of data. This is true in part, because these terms are closely related and sometimes used interchangeably.

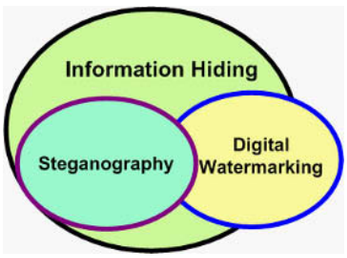
Information hiding is a general term that involves message embedding in some host media (Cox, Miller & Bloom, 2002). The purpose of information hiding is to make the information imperceptible or to keep the existence of the information secret. Steganography means "covered writing," a term derived from the Greek literature. Its purpose is to conceal the very existence of a message. Digital watermarking, however, embeds information into the host documents, but the embedded information may be visible (e.g., a company logo) or invisible (in which case, it is similar to steganography.)

Steganography and digital watermarking differ in several ways. First, the watermarked messages are related to the host documents (Cox et al., 2002). An example is the ownership information inserted into an image. Second, watermarks do not always have to be hidden. See Taylor, Foster and Pelly (2003) for applications of visible watermarks. However, visible watermarks are typically not considered steganography by definition (Johnson & Jajodia, 1998). Third, watermarking requires additional "robustness" in its algorithms. Robustness refers to the ability of a watermarking algorithm to resist removal or manipulation attempts (Craver, Perrig & Petitcolas, 2000; Acken, 1998). This characteristic deters attackers by forcing them to spend an unreasonable amount of computation time and/or by inflicting an unreasonable amount of damage to the watermarked documents in the attempts of watermark extraction. Figure 1 shows that there are considerable overlaps in the meaning and even the application of the three terms. Many of the algorithms in use today are in fact shared among information hiding, steganography and digital watermarking. The difference relies largely on "the intent of use" (Johnson & Jajodia, 1998). Therefore, discussions in the rest of the paper on watermarking also apply to steganography and information hiding, unless specifically mentioned otherwise.

To be consistent with the existing literature, a few terms are used in the rest of this article. *Cover work* refers to the host document (text, image, multimedia or other media content) that will be used to embed another document. This other document to be embedded is not limited to only text messages. It can be another image or other media content. *Watermark* refers to this latter document that will be

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Information hiding, steganography and digital watermarking



embedded in the cover work. The result of this embedding is called a *stego-object*.

CHARACTERISTICS OF EFFECTIVE WATERMARKING ALGORITHMS

Watermarking algorithms are not created equal. Some will not survive simple image processing operations, while others are robust enough to deter attackers from some forms of modifications. Effective and robust image watermarking algorithms should meet the following requirements:

- Modification tolerance. They must survive common document modifications and transformations (Berghel, 1997).
- Ease of authorized removal. They must be detectable and removable easily by authorized users (Berghel, 1997).
- Difficult unauthorized modifications. They also must be difficult enough to discourage unauthorized modifications.

In addition to the above requirements for image watermarking algorithms, Mintzer, Braudaway and Bell (1998) suggest the following for watermarking digital motion pictures:

- **Invisibility.** The presence of the watermark should not degrade the quality of motion pictures.
- **Unchanged compressibility.** The watermark should not affect the compressibility of the media content.
- **Low cost.** Watermark algorithms may be implemented in the hardware that only adds insignificant cost and complexity to the hardware manufacturers.

The main focus of these requirements concerns the capabilities of watermarking algorithms to survive various attacks or full/partial changes to the stego-object. However, the fundamental requirement for most algorithms is unobtrusiveness. Unless the goal of using an algorithm is to render the host medium unusable or partially unavailable, many algorithms will not produce something perceptibly different from the cover work. However, theoretically speaking, stego-objects are hardly the same as the cover work when something is embedded into the cover work.

When it comes to watermarking text documents, most of the above requirements apply. A text watermarking algorithm should not produce something that is easily detectable or render the resulting stego-object illegible. Different from many image or multimedia watermarking techniques, which produce imperceptible watermarks, text watermarking techniques typically render a visible difference if the cover work and stego-object are compared side by side.

DIGITAL WATERMARKS IN USE

Authentication of the host document is one important use of digital watermarks. In this scenario, a watermark is inserted into the cover work, resulting in a stego-object. Stripping off the watermark should yield the original cover work. Common uses of authentication watermarks include verification of object content (Mintzer, Braudaway & Bell, 1998) and copyright protection (Acken, 1998). The general concept of watermarking works in the following way.

$\boldsymbol{W} + \boldsymbol{M} = \boldsymbol{S}$, where $\boldsymbol{W}$ is the cover work, $\boldsymbol{M}$ is the watermark and $\boldsymbol{S}$ is the stego-object. The + operator embeds the watermark $\boldsymbol{M}$ into the cover work $\boldsymbol{W}$. (1.1)

The properties of watermarks used for authentication imply the following:

$$S - M' = W',\ W' \cong W \text{ and } M' \cong M, \quad (1.2)$$

where S is the stego-object, M' is the watermark to be stripped off from S, W' is the object with the M' stripped off. Theoretically, W' cannot be the same as W for watermarking algorithms that actually change W. However, invisible or imperceptible watermarks typically render an object that is perceived the same as the cover work in human eyes or ears. For this reason, the W' and W should be "perceived" as identical or similar. As (1.2) concerns watermarking for authentication, the main requirement is that the decoded watermark M' should be the same as the original watermark M for the authentication to work. In a more complex scenario similar to the concept of public key cryptography, a watermark can be considered as a "key" to lock or encrypt information, and another watermark will be used to unlock or decrypt the information. The two watermarks involved may bear little or no relationship with each other. Therefore, the $M' \cong M$ requirement may be relaxed for this scenario.

Watermarks can also be used in systems that require non-repudiation (Mintzer et al., 1998). Non-repudiation means a user cannot deny that something is created for him/her or by him/her. An example is that multiple copies of the cover work need to be distributed to multiple recipients. Before distribution, each copy is embedded with an identification watermark uniquely for the intended recipient. Unauthorized redistributions by a recipient can be easily traced since the watermark reveals the recipient's identity. This model implies the following:

$$W + M_{\{1, 2, ..., n\}} = S_{\{1, 2, ..., n\}}, \quad (1.3)$$

where M_1 is the watermark to be inserted into the copy for the first recipient, M_2 is for the second recipient and so on. S_1 is the stego-object sent to the first recipient and so on.

Generally, watermarks are expected to meet the robust requirements stated above, but in some cases, a "fragile" watermark is preferred. Nagra, Thomborson and Collberg (2002) suggest that software licensing could be enhanced with a licensing mark—a watermark that embeds information in software controlling how the software is used. In this scenario, a decryption key is used to unlock the software or grant use privileges. If the watermark is damaged, the decryption key should become ineffective; thus, the user is denied access to certain software functions or to the entire software. The fragility of the watermark in this example is considered more of a feature than a weakness.

Since the robust requirement is difficult to meet, some studies (e.g., Kwok, 2003) started to propose a model similar to digital certificates and certificate authorities in the domain of cryptography. A watermark clearance center is responsible for resolving watermarking issues, such as judging the ownership of a cover work. This approach aims at solving the deadlock problem where a pirate inserts his watermark in publicly available media and claims the ownership of such media.

CONCEALMENT IN DATA SOURCES

As information hiding, steganography and digital watermarking continue to attract research interests, the number of proposed algorithms mushrooms accordingly. It is difficult to give all algorithms a comprehensive assessment, due to the limited space in this article. Nonetheless, this section provides an overview of selected algorithms. The intent of this section is to offer a basic understanding of information hiding techniques.

Hiding Information in Text Documents

Information hiding techniques in plain text documents are very limited and susceptible to detection. Slight changes to a word or an extra punctuation symbol are noticeable to casual readers. With formatted text documents, the formatting styles add a wealth of options to information hiding techniques. Kankanhalli and Hau (2002) suggest the following watermarking techniques for electronic text documents:

- **Line shift encoding.** Vertical line spacing is changed to allow for message embedding. Each line shift may be used to encode one bit of data. This method works best in formatted text documents.
- **Word shift encoding.** Word spacing is changed to allow for message embedding. As with line shift encoding, word shift encoding is best suited in formatted text documents.

- **Feature encoding.** In formatted text documents, features and styles (such as font size, font type and color) may be manipulated to encode data.
- **Inter-character space encoding.** Spacing between characters is altered to embed data. This approach is most suited for human languages, such as Thai, where no large inter-character spaces are used.
- **High-resolution watermarking.** A text document is programmed to allow for resolution alteration so a message can be embedded.
- **Selective modifications of the document text.** Multiple copies of a master document are made, with modifications to a portion of the text. The text portion selected for modification is worded differently but with the same meaning so that each copy of the master document receives its own unique wording or word modifications in the selected text segments.
- Other embedding techniques to aid in encoding and decoding of watermarks.

Watermarking Images

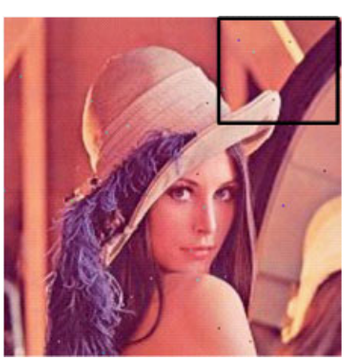
The simplest algorithm of image watermarking is the least significant bit (LSB) insertion. This approach replaces the LSBs of the three primary colors (i.e., red, green and blue) in those selected pixels with the watermark. To hide a single character in the LSBs

Figure 2. Text watermarked into an image

(a) Lena. Courtesy of the Signal and Image Processing Institute at the University of Southern California.

(b) Lena with the message "Digital watermarking is a fun topic" hidden.

(c) Figure 2(b) with selected pixels highlighted in color.



(d) The upper right corner of Figure 2(c) zoomed e00%. Colored dots are pixels with information embedded.

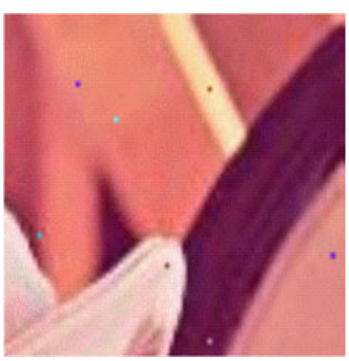


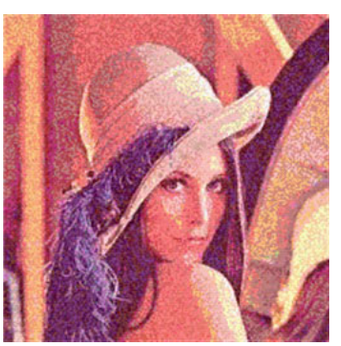
Figure 3. Watermark as an image—the extended Kurak-McHugh model

(a) Arctic Hare. Courtesy of Robert E. Barber, Barber Nature Photography. This image will be used as the watermark to be embedded in Figure 2(a).

(b) Extended Kurak-McHugh model. Lena with four MSBs of Figure 3(a) embedded.

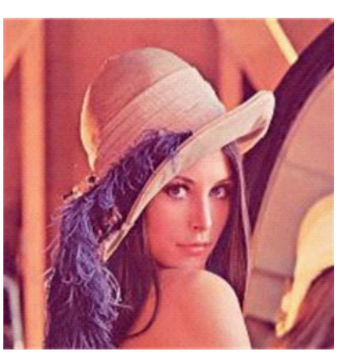
(c) Arctic Hare extracted from Figure 3(b).

(d) Lena with six MSBs of Figure 3(a) embedded.



(e) Arctic Hare extracted from Figure 3(d).

(f) Lena with two MSBs of Figure 3(a) embedded.



(g) Arctic Hare extracted from Figure 3(f).

of pixels in a 24-bit image, three pixels have to be selected. Since a 24-bit image uses a byte to represent each primary color of a pixel, each pixel then offers three LSBs available for embedding. A total of three pixels (9 LSBs) can be used to embed an 8-bit character, although one of the nine LSBs is not used. Figure 2 shows the LSB algorithm hiding the message "Digital watermarking is a fun topic" in an image. Even with the message hidden in the image, the stego-object is imperceptibly similar to the cover work. Figure 2(c) highlights those pixels that contain the text message and Figure 2(d) magnifies the upper-right corner of Figure 2(c) by 300%. The LSB insertion, although simple to implement, is vulnerable to slight image manipulation. Cropping, slicing, filtering and transformation may destroy the information hidden in the LSBs (Johnson & Jajodia, 1998).

The Kurak-McHugh model (Kurak & McHugh, 1992) offers a way to hide an image into another one. The main idea is that the *n* LSBs of each pixel in the cover work are replaced with the *n* most significant bits (MSBs) from the corresponding pixel of the watermark image. Figure 3 shows an extended version of the Kurak-McHugh model. The extension allows for embedding watermark MSBs into randomly selected pixels in the cover work. The figure also shows that the more MSBs are embedded, the coarser the resulting stego-object, and vice versa.

A similar approach to embed text messages in a cover work is proposed in Moskowitz, Longdon and Chang (2001). The eight bits of each character in a text message are paired. Each pair of bits is then embedded in the two LSBs of a randomly selected pixel. A null byte is inserted into the cover work to indicate the end of the embedded message.

The Patchwork algorithm (Bender, Gruhl, Morimoto & Lu, 1996) hides data in the difference of luminance between two "patches." The simplest form of the algorithm randomly selects a pair of pixels. The brightness of the first pixel in the pair is raised by a certain amount, and the brightness of the second pixel in the pair is lowered by another amount. This allows for embedding of "1," while the same process in reverse is used to embed a "0." This step continues until all bits of the watermark message are embedded. An extension of Patchwork includes treating patches of several points rather than single pixels. This algorithm is more robust to survive several image modifications, such as cropping, and gamma and tone scale correction.

Watermarking Other Forms of Media

Techniques for watermarking other types of media are also available in the literature. Bender et al. (1996) suggest several techniques for hiding data in audio files:

- **Low-bit encoding:** Analogues to the LSB approach in watermarking image files; the low-bit encoding technique replaces the LSB of each sampling point with the watermark.
- **Phase coding:** The phase of an initial audio segment is replaced with a reference phase that represents the data to be embedded. The phase of subsequent segments is adjusted to preserve the relative phase between segments.
- **Echo data hiding:** The data are hidden by varying the initial amplitude, decay rate and offset parameters of a covert work. Changes in these parameters introduce mostly inaudible/imperceptible distortions. This approach is similar to listening to music CDs through speakers where one listens to not just the music but also to the echoes caused by room acoustics (Gruhl, Lu & Bender, 1996). Therefore, the term "echo" is used for this approach.

Media that rely on internal cohesion to function properly have an additional constraint to watermarking algorithms. For example, software watermarking algorithms face an initial problem of location identification for watermarks. Unlike image files, an executable file offers very limited opportunities (e.g., some areas in the data segment) for watermarking. Interested readers of software watermarking should review Collberg and Thomborson (1999, 2002). Another example of media requiring internal cohesion is relational databases, in which case certain rules, such as database integrity constraints and requirements for appropriate keys, have to be maintained (Sion, Atallah & Prabhakar, 2003).

To increase the level of security, many watermarking algorithms involve selection of random pixels (or locations) to embed watermark bits. The process may begin with a carefully selected

password as the "seed" to initialize the random number generator (RNG). The RNG is then used to generate a series of random numbers (or locations) based on the seed. The decoding process works in a similar way, using the right password to initialize the RNG to locate the correct pixels/locations that have hidden information.

CONCLUSION

Information hiding, digital watermarking and steganography has received much interest in the last decade. Many of the algorithms strive for the tradeoff between the embedding capacity (bandwidth) and resistance to modification of the stego-objects (the robustness requirement). Algorithms with a high capacity of data concealment may be less robust, and vice versa. However, robust algorithms are not applicable nor needed for applications such as the licensing problem, in which fragility of watermarks are preferred. Furthermore, the whole set of watermarking techniques assume that a cover work can be modified without perceptible degradation or damage to the cover work. For digital contents that are less tolerable to even minor changes, information hiding techniques may not be the best solution.

Once an embedding technique is known, an attacker can easily retrieve the watermark; therefore, the goal of hiding fails. The information hiding community also recommended bridging steganography with cryptography (Anderson & Petitcolas, 1998). In this combination, a watermark message is encrypted before being embedded into a cover work. However, this also introduces the computation speed vs. key distribution tradeoffs currently present in cryptographic algorithms. Generally, secret key cryptographic algorithms, although faster to compute, require that the key be distributed securely. Conversely, public key cryptographic algorithms impose a longer computation time, but ease the key distribution problem. Information hiding will undoubtedly attract more research. Existing algorithms will be refined and new algorithms will emerge to improve resistance from digital modifications.

REFERENCES

Acken, J.M. (1998). How watermarking adds value to digital content. *Communications of the ACM*, *41*(7), 74-77.

Anderson, R.J., & Petitcolas, F.A.P. (1998). On the limits of steganography. *IEEE Journal of Selected Areas in Communications*, *16*(4), 474-481.

Bender, W., Gruhl, D., Morimoto, N., & Lu, A. (1996). Techniques for data hiding. *IBM Systems Journal*, *35*(3&4), 313-336.

Berghel, H. (1997). Watermarking cyberspace. *Communications of the ACM*, *40*(11), 19-24.

Collberg, C., & Thomborson, C. (1999). Software watermarking: models and dynamic embeddings. *Proceedings of the 26th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, (pp. 311-324).

Collberg, C., & Thomborson, C. (2002). Watermarking, tamper-proofing, and obfuscation – Tools for software protection. *IEEE Transactions on Software Engineering*, *28*(8), 735-746.

Cox, I.J., Miller, M.L., & Bloom, J.A. (2002). *Digital watermarking*. London: Academic Press.

Craver, S., Perrig, A., & Petitcolas, F.A.P. (2000). Robustness of copyright marking systems. In S. Katzenbeisser & F.A.P. Petitcolas (Eds.), *Information hiding: Techniques for steganography and digital watermarking* (pp. 149-174). Norwood: Artech House.

Gruhl, D., Lu, A., & Bender, W. (1996). Echo data hiding. *International Workshop on Information Hiding*, 295-315.

Johnson, N.F., & Jajodia, S. (1998). Exploring steganography: Seeing the unseen. *IEEE Computer*, February, 26-34.

Kankanhalli, M.S., & Hau, K.F. (2002, Jan/April). Watermarking of electronic text documents. *Electronic Commerce Research*, *2*(1-2), 169-187.

Kurak, C., & McHugh, J. (1992). A cautionary note on image downgrading. *Computer Security Applications Conference*, San Antonio, Texas, (pp. 153-159).

Kwok, S.H. (2003). Watermark-based copyright protection system security. *Communications of the ACM*, *46*(10), 98-101.

Mintzer, F., Braudaway, G.W., & Bell, A.E. (1998). Opportunities for watermarking standards. *Communications of the ACM*, *41*(7), 56-64.

Nagra, J., Thomborson, C., & Collberg, C. (2002). A functional taxonomy for software watermarking. *Proceedings of the 25th Australasian conference on Computer Science*, (vol. 4, pp. 177-186).

Sion, R., Atallah, M., & Prabhakar, S. (2003, June 9-12). Rights protection for relational data. *Proceedings of the 2003 ACM SIGMOD,* San Diego, California.

Taylor, A., Foster, R., & Pelly, J. (2003). Visible watermarking for content protection. *SMPTE Mostion Imaging*, Feb/March, 81-89.

KEY TERMS

Cover Work: The host media in which a message is to be inserted or embedded.

Cryptography: A study of making a message secure through encryption. Secret key and public key are the two major camps of cryptographic algorithms. In secret key cryptography, one key is used for both encryption and decryption; in public key cryptography, two keys (public and private) are used.

Digital Watermarking: This concerns the act of inserting a message into a cover work. The resulting stego-object can be visible or invisible.

Information Hiding: The umbrella term referring to techniques of hiding various forms of messages into cover work.

Least Significant Bit (LSB): An LSB refers to the last or the right-most bit in a binary number. The reason it is called LSB is because changing its value will not dramatically affect the resulting number.

Steganography: Covered writing. It is a study of concealing the existence of a message.

Stego-Object: This is the cover work with a watermark inserted or embedded.

Information Security Management

Mariana Hentea
Southwestern Oklahoma State University, USA

INFORMATION SECURITY MANAGEMENT OVERVIEW

Information security management is the framework for ensuring the effectiveness of information security controls over information resources to ensure no repudiation, authenticity, confidentiality, integrity and availability of the information. Organizations need a systematic approach for information security management that addresses security consistently at every level. However, the security infrastructure of most organizations came about through necessity rather than planning, a reactive-based approach as opposed to a proactive approach (Gordon, Loeb & Lucyshyn, 2003). Intrusion detection systems, firewalls, anti-virus software, virtual private networks, encryption and biometrics are security technologies in use today. Many devices and systems generate hundreds of events and report various problems or symptoms. Also, these devices may all come at different times and from different vendors, with different reporting and management capabilities and—perhaps worst of all—different update schedules. The security technologies are not integrated, and each technology provides the information in its own format and meaning. In addition, these systems across versions, product lines and vendors may provide little or no consistent characterization of events that represent the same symptom. Also, the systems are not efficient and scalable because they rely on human expertise to analyze periodically the data collected with all these systems. Network administrators regularly have to query different databases for new vulnerabilities and apply patches to their systems to avoid attacks. Quite often, different security staff is responsible and dedicated for the monitoring and analysis of data provided by a single system. Security staff does not periodically analyze the data and does not timely communicate analysis reports to other staff. The tools employed have very little impact on security prevention, because these systems lack the capability to generalize, learn and adapt in time.

Therefore, the limitations of each security technology combined with attacks growth impact the efficiency of information security management and increase the activities to be performed by network administrators. Specific issues include data collection, data reduction, data normalization, event correlation, behavior classification, reporting and response.

Cyber security plans call for more specific requirements for computer and network security as well as emphasis on the availability of commercial automated auditing and reporting mechanisms and promotion of products for security assessments and threat management (Hwang, Tzeng & Tsai, 2003; Chan, 2003; Leighton, 2004). Recent initiatives to secure cyberspace are based on the introduction of cyber-security priorities that call for the establishment of information sharing and analysis centers. Sharing information via Web services brings benefits as well as risks (Dornan, 2003). Security must be considered at all points and for each user. End-to-end security is a horizontal process built on top of multiple network layers that may have security or no security. Security is a process based on interdisciplinary techniques (Mena, 2004; Maiwald, 2004).

The following sections discuss security threats impact, emerging security management technologies, information security management solutions and security event management model requirements.

SECURITY THREATS IMPACT

Information security means protecting information and systems from security threats such as unauthorized access, use, disclosure, disruption, modification or destruction of information. The frequency of information security breaches is growing and common among most organizations. Internet connection is increasingly cited as a frequent point of attack and

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

likely sources of attacks are independent hackers and disgruntled employees. Despite the existence of firewalls and intrusion detection systems, network administrators must decide how to protect systems from malicious attacks and inadvertent cascading failures. Effective management of information security requires understanding the processes of discovery and exploitation used for attacking. An attack is the act of exploiting a vulnerability that is a weakness or a problem in software (a bug in the source code or flaw in design). Software exploits follow a few patterns; one example is buffer overflow. An attack pattern is defined as a "blueprint for creating a kind of attack" (Hoglund & McGraw, 2004, p. 26). Buffer overflow attacks follow several standard patterns, but they may differ in timing, resources used, techniques and so forth.

Broad categories of attack patterns include network scanning, operating system stack identification, port scans, traceroute and zone transfers, target components, choosing attack patterns, leveraging faults in the environment, using indirection and planting backdoors. Typically, an attack is a set of steps. The first phase is discovery or network reconnaissance. The attacker collects information about the target using public databases and documents as well as more invasive scanners and grabbers. Then, the attacker tries to discover vulnerabilities in the services identified, either through more research or by using a tool designed to determine if the service is susceptible. From a damage point of view, scans typically are harmless. Intrusion detection systems classify scans as low-level attacks because they don't harm servers or services. However, scans are precursors to attacks. If a port is discovered open, there is no guarantee that the attacker will not return, but it is more likely that he will and the attack phase begins. Several services and applications are targets for attack.

"Web within Web" (Castro-Leon, 2004, p. 42) or Web services such as UDDI (finding a Web site), WSDL (site description), SOAP (transport protocol) and XML (data format) are security concerns. Much Web services security technology is still being developed and has not stabilized enough to inspire confidence. For example, protocols (SOAP) are lacking security, or specifications for Web services security (WS-SEC) are still evolving, and providing security in hardware is not an option because the specifications are not ready to be set in silicon (Dornan, 2003). On the other hand, standards themselves do not guarantee interoperability or security. It depends on how vendors implement the standards (Navas, 2002). Sometimes, Web security requires use of public key infrastructure (PKI). However, PKI is complex and has been a difficult infrastructure to manage, and the cost of managing has been detrimental to many organizations (Geer, 2003). Also, PKI infrastructure is not readily available in many parts of the world.

Spam is another threat that is increasing each year. The best anti-spam solutions rely on a set of detection methods such as heuristics, white and black lists, and signature matching. Choosing the right solution for an organization implies understanding how common spam filters operate, and what their tradeoffs are. Filtering the spam requires human intervention even when tools are available. Bayesian filtering promises a future where most of the spam could be detected and blocked automatically, but these tools are too complex for a mass audience, and wide-scale adoption is probably a few years out (Conry-Murray, 2003).

A very common threat is unauthorized access. This can be prevented via access controls enhanced with biometric systems, a type of access control mechanism used to verify an individual's identity. Biometric systems fall into two categories: authentication and identification, with authentication systems by far more common. Authentication systems are reliable and efficient if the subject base is small and the biometric readers are accurate and durable. A database with biometric data presents a natural target for theft and malicious and fraudulent use (Johnson, 2004). Voice authorization products are becoming popular because they allow remote authentication (Vaughan-Nichols, 2004), but the technology is the least accurate and network administrators have to use it cautiously until researchers improve it.

Moving data over back-end networks, remote locations, shared recovery centers and outsourced information technology facilities also expose information to threats (Hughes & Cole, 2003). The next section describes major trends in information security management.

EMERGING SECURITY TECHNOLOGIES

Surveys of security technologies indicate that most organizations use security technologies such as firewalls, anti-virus software, some kind of physical security to protect their computer and information assets or some measures of access control (Richardson, 2003). Technologies such as virtual private networks (Zeng & Ansari, 2003) and biometrics using a fingerprint are predicted to grow very fast, and others are still emerging. The newest version of an intrusion detection system based on open-source Snort 2.0 supports a high-performance multi-pattern search engine with an anti-denial of service strategy (Norton & Roelker, 2003). However, detecting distributed denial-of-service (DDoS) is still emerging due to the complexity of technical problems not known to build defenses against this type of attack. Current technologies are not efficient for large-scale attacks, and comprehensive solutions should include attack prevention and preemption, attack detection and filtering, and attack source trace back and identification (Chang, 2002).

In addition, new protocols are defined and old protocols are enhanced. One example is IP security protocol (IPSec) defined by IETF. IPSec protocol is implemented for new IPv6 services in the very high-broadband-speed networks for new-generation Internet applications (Adam, Fillinger, Astic, Lahmadi & Brigant, 2004). In the near future, the network environment is expected to include hosts that support IPv4 and IPv6 protocols (Tatipamula, Grosette & Esaki, 2004), and new tools are needed for network administrators.

Other trends include integration of information security with physical security (Hamilton, 2003), self-securing devices and sensor networks. Self-securing devices offer new capabilities for dealing with intrusions, such as preventing undetectable tampering and deletion. If the detection mechanism discovers a change, an alert is sent to the network administrator for action (Cummings, 2002). Sensor networks are essential to the creation of smart spaces, which embed information technology in everyday home and work environments (Marculescu, Marculescu, Sungmee & Jayraman, 2003; Ashok & Agrawal, 2003). The privacy and security issues posed by sensor networks and sensor detectors represent a rich field of research problems (Chan & Perrig, 2003).

Within the past years, a new security market has emerged, known as Security Event Management (SEM), which is part of Security Incident Management. SEM includes the processes that an organization uses to ensure the collection, security and analysis of security events as well as notification and response to security events. Although limited on capabilities, new products based on solutions for SEM are emerging slowly. The new products lack the prevention capability and still rely on human expertise to make decisions, or require substantial manual configurations up front. Data mining and other techniques for extracting coherent patterns of information from a call are near the top of the research agenda. For example, focusing on telephone calls from a particular installation, searching for specific words and phrases in e-mails, or using voice recognition techniques all are deployed. Cell and satellite phones can also reveal a caller's location (Wallich, 2003). The following section discusses issues and solutions for information security management.

INFORMATION SECURITY MANAGEMENT SOLUTIONS

IBM's manifesto (Kephart & Chess, 2003) points out difficulties in managing computing systems because their complexity is approaching the limits of human capability while there is need for increased interconnectivity and integration. Systems are becoming too complex for even the most skilled system integrators to install, configure, optimize and maintain. Information security management is no exception. One proposed solution is autonomic computing – computing systems that can manage themselves given high-level objectives from administrators. These systems require capabilities for self-configuration, self-optimization, self-healing and self-protection. Therefore, the success of autonomic computing is in the future, many years ahead.

In more sophisticated autonomic systems, machine learning by a single agent is not sufficient, and multi-agent solutions are proposed, although there are no guarantees of convergence because agents are adapting to one another. The agents change

their behavior, making other agents change their behavior. Artificial intelligence (AI) techniques enhance agent capabilities. Intelligent agents and multiagent systems are among the most growing areas of research and development. Intelligent agent technology is not a single, new technology, but rather the integrated application of technologies such as network, Internet and AI techniques. Learning in multiagent systems is a challenging problem, so it is optimization. Intelligent models of large networked systems will let autonomic elements or systems detect or predict overall performance problems from a stream of sensor data from individual devices. At long time scales—during which the configuration of the system changes—new methods will be feasible to automate the aggregation of statistical variables to reduce the dimensionality of the problem to a size amenable to adaptive learning and optimization techniques that operate on shorter time scales.

Contrary to autonomous systems, new systems focus on human-agent effective interaction such that security policies can control agent execution and communicate with a human to ensure that agent behavior conforms to desired constraints and objectives of the security policies (Bradshaw, Cabri & Montanari, 2003; Bhatti, Bertino, Ghafoor & Joshi, 2004). A Microsoft project on next-generation secure-computing base is focused on building robust access control while retaining the openness of personal computers by providing mechanisms that allow operating systems and applications to protect themselves against other software running on the same machine (England, Lampson, Manferdelli, Peinado & Williams, 2003). Still, robustness against software attacks will depend on hardware and software free from security relevant bugs. A business solution is to enforce quality security to software manufacturers and liability to the computer industry (Schneir, 2004).

Efficient information security management requires an SEM approach with enhanced real-time capabilities, adaptation and generalization to predict possible attacks and to support humans' actions. The following section discusses major requirements for the SEM model.

SEM MODEL REQUIREMENTS

The objective of the SEM is the real-time analysis and correlation of events. The model should be adaptable and capable to support monitoring and control of the network to include data collected by all security technologies and network management systems instead of relying on data provided by each single system. Although advanced techniques based on AI are emerging, these are still focused on a limited scope. For example, Sun Microsystems developed a host-based intrusion detection system using expert systems techniques for the Sun Solaris platform (Lidqvist & Porras, 2001). The SEM model should be cost effective such that organizations could afford the use of advanced technologies for security protection (Wallich, 2003).

The SEM model should be a hybrid model based on the integration of traditional statistical methods and various AI techniques to support a general system that operates automatically, adaptively and proactively (Hentea, 2003, 2004). Statistical methods have been used for building intrusion and fault detection models (Manikopoulos & Papavassiliou, 2002). AI techniques such as data mining, artificial neural networks, expert systems and knowledge discovery can be used for classification, detection and prediction of possible attacks or ongoing attacks. Machine learning technique is concerned with writing programs that can learn and adapt in real time. This means that the computer makes a prediction and then, based on the feedback as to whether it is correct, learns from this feedback. It learns through examples, domain knowledge and feedback. When a similar situation arises in the future, the feedback is used to make the same prediction.

The security model should include identification and selection of data needed to support useful feedback to a network administrator or security staff. In addition, the type of feedback available is important. Direct feedback entails specific information about the results and impact of each possible feedback. Indirect feedback is at a higher level, with no specific information about individual change or

predictions but whether the learning program can propose new strategies and changes. Another important factor to consider is that systems, software and security policies change themselves over time and across different platforms and businesses. These special circumstances have to be included in the machine learning program to support the user and the security management process. In addition, the machine learning program should support a knowledge base to enrich the learning environment that allows the user to answer about unknowns in the system.

CONCLUSION

Security event management solutions are needed to integrate threat data from various security and network products to discard false alarms, correlate events from multiple sources and identify significant events to reduce unmanaged risks and improve operational security efficiency. There is a need for increased use of automated tools to predict the occurrence of security attacks. Auditing and intelligent reporting mechanisms must support security assessment and threat management at a larger scale and in correlation with the past, current and future events.

REFERENCES

Adam, Y., Fillinger, B., Astic, I., Lahmadi, A., & Brigant, P. (2004). Deployment and test of IPv6 services in the VTHD network. *IEEE Communications Magazine, 42*(1), 98-104.

Ashok, R.L., & Agrawal, D.P. (2003). Next-generation wearable networks. *IEEE Computer, 36*(11), 31-39.

Bhatti, R., Bertino, E., Ghafoor, A., & Joshi, J.B.D. (2004). XML-based specification for Web services document security. *IEEE Computer, 37*(4), 41-49.

Bradshaw, J.M. Cabri, J., & Montanari, R. (2003). Taking back cyberspace. *IEEE Computer, 36*(7), 89-92.

Castro-Leon, E. (2004). The WEB within the WEB. *IEEE Spectrum, 41*(2), 42-46.

Chan, H., & Perrig, A. (2003). Security and privacy in sensor networks. *IEEE Computer, 36*(10), 103-105.

Chang, R.K.C. (2002). Defending against flooding-based distributed denial-of-service attacks: A tutorial. *IEEE Communications Magazine, 40*(10), 42-51.

Conry-Murray, A. (2003). Fighting the spam monster – and winning. *Network Magazine, 18*(4), 24-29.

Cummings, R. (2002). The evolution of information assurance. *IEEE Computer, 35*(12), 65-72.

Dornan, A. (2003). XML: The end of security through obscurity? *Network Magazine, 18*(4), 36-40.

England, P., Lampson, B., Manferdelli, J., Peinado, M., & Williams, B. (2003). A trusted open platform. *IEEE Computer, 36*(7), 55-62.

Geer, D. (2003). Risk management is still where the money is. *IEEE Computer, 36*(12), 129-131.

Gordon, L.A., Loeb, M.P., & Lucyshyn, W. (2003). Information security expenditures and real options: A wait-and-see approach. *Computer Security Journal, XIX*(2), 1-7.

Hamilton, C. (2003). Holistic security. *Computer Security Journal, XIX*(1), 35-40.

Hentea, M. (2003). Intelligent model for cyber attack detection and prevention. *Proceedings of the ISCA 12th International Conference Intelligent and Adaptive Systems and Software Engineering* (pp. 5-10).

Hentea, M. (2004). Data mining descriptive model for intrusion detection systems. *Proceedings of the 2004 Information Resources Management Association International Conference,* 1118-1119. Hershey, PA: Idea Group Publishing.

Hoglund, G., & McGraw, G. (2004). Attack patterns. *Computer Security Journal, XIX*(2), 15-32.

Hughes, J., & Cole, J. (2003). Security in storage. *IEEE Computer, 36*(1), 124-125.

Hwang, M3-S., Tzeng, S-F., & Tsai, C-S. (2003). A new secure generalization of threshold signature scheme. *Proceedings of International Technology for Research and Education* (pp. 282-285).

Johnson, M.L. (2004). Biometrics and the threat to civil liberties. *IEEE Computer, 37*(4), 90-92.

Kephart, J.O., & Chess, D.M. (2003). The vision of automatic computing. *IEEE Computer, 36*(1), 41-50.

Leighton, F.T. (2004). Hearing on "The state of cyber security in the United States government." *Computer Security Journal, XX*(1), 15-22.

Lidqvist, U., & Porras, P.A. (2001). eXpert-BSM: A host-based intrusion detection solution for Sun Solaris. *Proceedings of the 17th Annual Computer Security Applications Conference* (pp. 240-251).

Maiwald, E. (2004). *Fundamentals of network security.* New York: McGraw Hill.

Manikopoulos, C., & Papavassiliou, S. (2002). Network intrusion and fault detection: A statistical anomaly approach. *IEEE Communications Magazine, 40*(10), 76-82.

Marculescu, D., Marculescu, R., Sungmee, P., & Jayraman, S. (2003). Ready to ware. *IEEE Spectrum, 40*(10), 29-32.

Mena, J. (2004). HOMELAND SECURITY Connecting the DOTS. *Software Development, 12*(5), 34-41.

Navas, D. (2002). What's next in integration: Manufacturing taps the Web for collaboration. *Supply&Chain Systems Magazine, 22*(9), 22-30, 56.

Norton, M., & Roelker, D. (2003). The new Snort. *Computer Security Journal, XIX*(1), 37-47.

Richardson, R. (2003). 2003 CSI/FBI computer crime and security survey. *Computer Security Journal, XIX*(2), 21-40.

Schneir, B. (2004). Hacking the business climate for network security. *IEEE Computer, 37*(4), 87-89.

Tatipamula, M., Grosette, P., & Esaki, H. (2004). IPv6 integration and coexistence strategies for next-generation networks. *IEEE Communications Magazine, 42*(1), 88-96.

Vaughan-Nichols, S. (2004). Voice authentication speaks to the marketplace. *IEEE Computer, 37*(3), 13-15.

Wallich, P. (2003). Getting the message. *IEEE Spectrum, 40*(4), 39-42.

Zeng, J., & Ansari, N. (2003). Toward IP virtual private network quality of service: A service provider perspective. *IEEE Communications Magazine, 41*(4), 113-119.

KEY TERMS

Artificial Neural Networks: Approach based on neural structure of the brain with the capability to identify and learn patterns from different situations as well as to predict new situations.

Data Mining: Approach for extracting coherent patterns of information from huge amounts of data and events.

Expert Systems: Approach designed to mimic human logic to solve complex problems.

Information Security Management: A framework for ensuring the effectiveness of information security controls over information resources; it addresses monitoring and control of security issues related to security policy compliance, technologies and actions based on decisions made by humans.

Intelligent Agent Technology: Integration of network, Internet and Artificial Intelligence techniques.

Security Event Management (SEM): An approach for the event detection, correlation and prevention of attacks, including automatic and automated enforcement of security policies.

Security Policy: Guidelines for security of the information, computer systems and network equipment.

Information Security Management in Picture Archiving and Communication Systems for the Healthcare Industry

Carrison KS Tong
Pamela Youde Nethersole Eastern Hospital and Tseung Kwan O Hospital, Hong Kong

Eric TT Wong
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

Like other information systems in banking and commercial companies, information security is also an important issue in the healthcare industry. It is a common problem to have security incidences in an information system. Such security incidences include physical attacks, viruses, intrusions, and hacking. For instance, in the U.S.A., more than 10 million security incidences occurred in the year of 2003. The total loss was over $2 billion. In the healthcare industry, damages caused by security incidences could not be measured only by monetary cost. The trouble with inaccurate information in healthcare systems is that it is possible that someone might believe it and do something that might damage the patient. In a security event in which an unauthorized modification to the drug regime system at Arrowe Park Hospital proved to be a deliberate modification, the perpetrator received a jail sentence under the Computer Misuse Act of 1990. In another security event (The Institute of Physics and Engineering in Medicine, 2003), six patients received severe overdoses of radiation while being treated for cancer on a computerized medical linear accelerator between June 1985 and January 1987. Owing to the misuse of untested software in the control, the patients received radiation doses of about 25,000 rads while the normal therapeutic dose is 200 rads. Some of the patients reported immediate symptoms of burning and electric shock. Two died shortly afterward and others suffered scarring and permanent disability.

BS7799 is an information-security-management standard developed by the British Standards Institution (BSI) for an information-security-management system (ISMS). The first part of BS7799, which is the code of practice for information security, was later adopted by the International Organization for Standardization (ISO) as ISO17799. The second part of BS7799 states the specification for ISMS. The picture-archiving and -communication system (PACS; Huang, 2004) is a clinical information system tailored for the management of radiological and other medical images for patient care in hospitals and clinics. It was the first time in the world to implement both standards to a clinical information system for the improvement of data security.

BACKGROUND

Information security is the prevention of, and recovery from, unauthorized or undesirable destruction, modification, disclosure, or use of information and information resources, whether accidental or intentional. A more proactive definition is the preservation of the confidentiality, integrity, and availability (CIA) of information and information resources. Confidentiality means that the information should only be disclosed to a selected group, either because of its sensitivity or its technical nature. Information integrity is defined as the assurance that the information used in making business decisions is created and maintained with appropriate controls to ensure that the information is correct, auditable, and reproducible. As far as information availability is concerned, information is said to be available when employees who are authorized access, and whose jobs require access, to the information can do so in a cost-effective manner that does not jeopardize the value of the information. Also, information must be consistently available to conduct business smoothly. Business-continuity planning (BCP) includes provisions for assuring the availability of the key resources (information, people, physical assets, tools, etc.) necessary to support the business function.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

The origin of ISO17799/BS7799 goes back to the days of the UK Department of Trade and Industry's (DTI) Commercial Computer Security Centre (CCSC). Founded in May 1987, the CCSC had two major tasks. The first was to help vendors of IT security products by establishing a set of internationally recognised security-evaluation criteria and an associated evaluation and certification scheme. This ultimately gave rise to the information technology security-evaluation criteria (ITSEC) and the establishment of the UK ITSEC scheme. The second task was to help users by producing a code of good security practices and resulted in the *Users Code of Practice* that was published in 1989. This was further developed by the National Computing Centre (NCC) and later a consortium of users, primarily drawn from British industry, to ensure that the code was both meaningful and practical from a user's point of view. The final result was first published as the British Standards guidance document PD 0003, *A Code of Practice for Information Security Management*, and following a period of further public consultation, it was recast as British Standard BS7799: 1995. A second part, BS7799-2: 1998, was added in February 1998. Following an extensive revision and public consultation period in 1997, the first revision of the standard, BS7799: 1999, was published in April 1999. Part 1 of the standard was proposed as an ISO standard via the "fast track" mechanism in October 1999, and then published with minor amendments as ISO/IEC 17799: 2000 on December 1, 2000. BS7799-2: 2002 was officially launched on September 5, 2002.

PACS is a filmless (Dreyer, Mehta, & Thrall, 2001) and computerized method of communicating and storing medical image data such as computed radiographic (CR), digital radiographic (DR), computed tomographic (CT), ultrasound (US), fluoroscopic (RF), magnetic resonance (MRI), and other special X-ray (XA) images. A PACS consists of image and data acquisition and storage, and display stations integrated by various digital networks. Full PACS handles images from various modalities. Small-scale systems that handle images from a single modality (usually connected to a single acquisition device) are sometimes called *mini-PACS*.

The medical images are stored in an independent format. The most common format for image storage is DICOM (Digital Imaging and Communications in Medicine), developed by the American College of Radiology and the National Electrical Manufacturers' Association.

Tseung Kwan O Hospital (TKOH) is a newly built general acute hospital (built in 1999) with 458 inpatient beds and 140 day beds. The hospital is composed of several clinical departments including medicine; surgery; paediatrics and adolescent medicine; eye, ear, nose, and throat; accident and emergency, and radiology. A PACS was built in its radiology department in 1999. The PACS was connected with the CR, CT, US, RF, DSA (Digital Subtraction Angiogram), and MRI system in the hospital. The hospital has become filmless since a major upgrade of the PACS in 2003.

An ISO17799/BS7799 ISMS was implemented in the TKOH PACS in 2003. During the implementation, a PACS security forum was established with the active participation of radiologists, radiographers, medical physicists, technicians, clinicians, and employees from the information technology department (ITD). After a BS7799 audit conducted in the beginning of 2004, the TKOH PACS was the world's first system with the ISMS certification.

In this article, the practical experience of the ISO17799/BS7799 implementation and the quality-improvement process of such a clinical information system will be explained.

MAIN FOCUS OF THE ARTICLE

In TKOH, the PACS serves the whole hospital including all clinical departments. The implementation of ISO17799 and BS7799 was started with the establishment of an ISMS for the PACS at the beginning of 2003. For effective implementation of ISO17799 and BS7799 in general, four steps will be required.

1. Define the scope of the ISMS in the PACS.
2. Make a risk analysis of the PACS.
3. Created plans as needed to ensure that the necessary improvements are implemented to move the PACS as a whole forward toward the BS7799 objective.
4. Consider other methods of simplifying the above and achieving compliance with minimum effect.

Implementation of BS7799 Controls in the TKOH PACS Security Forum

A PACS security forum was established for the effective management of all PACS-related security issues in the hospital. The members of the forum were the hospital chief executive, radiologist, clinician, radiographers, medical physicists, technicians, and representatives from the information technology department. One of the major functions of the PACS security forum was to make the security policies for the management of the PACS (Peltier, 2001a). Regular review of the effectiveness of the management was also required.

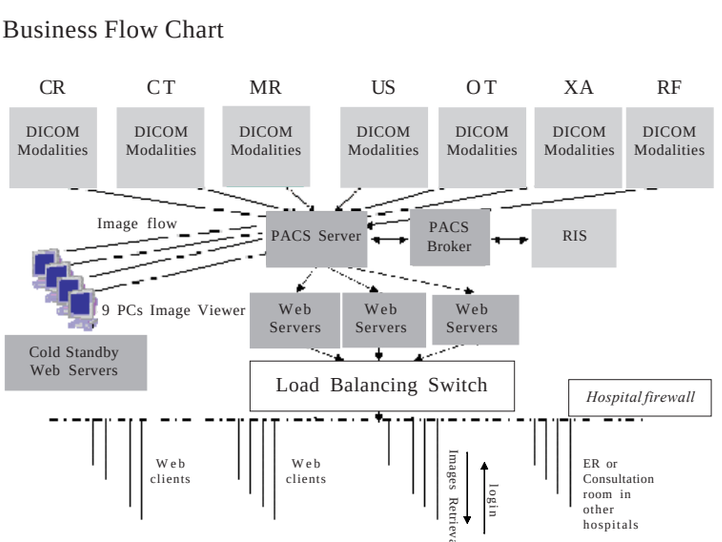
Business-Continuity Plan

BCP (Calder & Watkins, 2003) is a plan that consists of a set of activities aimed at reducing the likelihood and limiting the impact of disaster events on critical business processes. By the practice of BCP, the impact and downtime of the hospital's PACS system operation due to some change or failure in the company operation procedure is reduced. BCP is used to make sure that the critical part of the PACS system operation is not affected by a critical failure or disaster. The design of this BCP is based on the assumption that the largest disaster is a complete breakdown of the PACS room in the radiology department of TKOH. The wards, the specialist outpatient department (SOPD), and the imaging modalities should still all be functional.

During the design of a BCP, a business-impact analysis (BIA) of the PACS was studied. The BIA was a study of the vulnerabilities of the business flow of the PACS, and it is shown in the following business flowchart.

In the above flowchart, image data were acquired by the CR, DR, CT, US, RF, MRI, XA, and other (OT) imaging modalities such as a film digitizer. The acquired image data were centrally archived to the PACS server, which connected to a PACS broker for the verification of patient demographic data with the information from the Radiology Information System (RIS). In the PACS server, a storage-area network (SAN), a magneto-optical disk (MOD) jukebox, and a tape library were installed for short-term, long-term, and backup storage. The updated or verified image was redirected to the Web server cluster (Menasce & Almeida, 2001) for image distribution to the entire hospital including the emergency room (ER) and consultation room. The load-balancing switch was used for nonstop service of image distribution to the clinicians. A cluster of Cisco

Figure 1. Business flowchart of the Tseung Kwan O Hospital picture-archiving and -communication system



switches was installed and configured for automatic fail-over and firewall purposes. The switches connecting between the PACS network and hospital network were maintained by the information technology department (*A Practical Guide to IT Security for Everyone Working in Hospital Authority*, 2004; *Security Operations Handbook*, 2004). A remote-access server was connected to the PACS for the remote service of the vendor.

Business-Impact Analysis

In the BIA (Peltier, 2001b), according to the PACS operation procedure, all potential risks and impacts were identified. The responsibilities of relevant teams or personnel were identified according to the business flow of the PACS. The critical risk(s), which may affect the business operation of the PACS, could be determined by performing a risk evaluation of the potential impact. One of the methods in the BIA was to consider the contribution of the possibility of risk occurrence for prioritization purposes. The result of the BIA is shown in the following table.

In table 1, the responsible person for each business subprocess was identified to be PACS team, radiologists, radiographers, clinicians, or the information technology department. The most critical subprocess in the TKOH PACS was associated with the Web servers. Once the critical subprocess was identified, the BCP could be designed for the system as shown in the following figure. A responsible person for the BCP was also assigned.

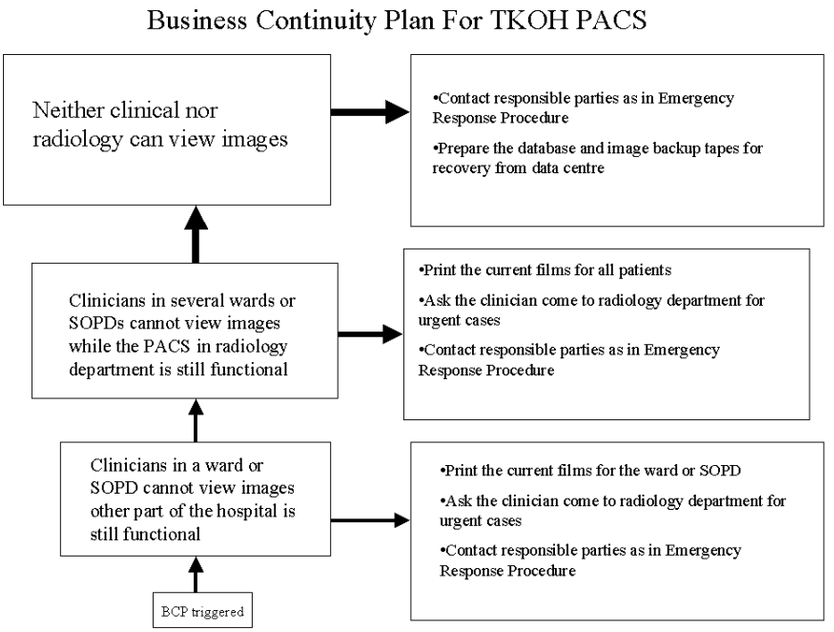
Disaster-Recovery Plan

Disaster-recovery planning (DRP; Toigo, 1996), as defined here, is the recovery of a system from a specific unplanned domain of disaster events such as natural disasters, or the complete destruction of the system. Following is the DRP for the TKOH PACS, which was also designed based on the result of the above BIA.

Table 1. Result of BIA

Process No.	Process Location	Risk	Subprocess	Responsible Person	Impact	Impact Level	Probability	Level of Importance
1	PACS broker	Hardware failure	Patient demographic-data retrieval	Radiographers, ITD	Manual input of patient demographic data	1	1	1
2	PACS servers	Hardware failure	Image receiving	PACS team	PACS cannot receive new images	2	1	2
3	SAN	Hardware failure	Image online storage	PACS team	No online image available in PACS. User still can view the images in the Web servers.	2	1	2
4	PACS servers	Hardware failure	Image verification	PACS team	Image data maybe different from what is in the RIS	1	1	2
5	Image viewers	Hardware failure	Image reporting	PACS team, radiographers	Radiologists cannot view images in the PACS server for advanced image processing and reporting. However, they can still see the images in the Web servers.	2	1	2
6	Jukebox	Hardware failure	Image archiving to MOD jukebox	PACS team	Long-term archiving of the images. There is a risk of lost images in the SAN.	2	2	4
7	Tape library	Hardware failure	Image archiving to tape library	PACS team	Another copy of long-term archiving. There is a risk of lost images in the SAN.	2	1	2
8	Jukebox	Hardware failure	Image prefetching from MOD jukebox	Radiologists, radiographers	Users cannot see the previous images. They cannot compare the present study with the previous.	2	2	4
9	Tape library	Hardware failure	Image prefetching from tape library	Radiologists, radiographers	Users cannot see the previous images. They cannot compare the present study with the previous.	2	1	2
10	Web servers	Hardware failure	Image distribution to clinicians	Clinicians, radiographers	The clinician cannot make a diagnosis without the images.	3	2	6
11	Cisco switches	Hardware failure	Image distribution through Cisco switches	Clinicians, radiographers	The clinician cannot make a diagnosis without the images.	3	1	3
12	Load-balancing switch	Hardware failure	Web-server load balancing	PACS team	The clinician cannot make a diagnosis without the images.	2	1	2
13	RAS server and Cisco router	System malfunction	Remote maintenance	PACS team	Vendor cannot do maintenance remotely.	1	1	1

Figure 2. Business-continuity plan for the TKOH PACS


Business Continuity Plan For TKOH PACS
Neither clinical nor radiology can view images
•Contact responsible parties as in Emergency Response Procedure
•Prepare the database and image backup tapes for recovery from data centre
Clinicians in several wards or SOPDs cannot view images while the PACS in radiology department is still functional
•Print the current films for all patients
•Ask the clinician come to radiology department for urgent cases
•Contact responsible parties as in Emergency Response Procedure
Clinicians in a ward or SOPD cannot view images other part of the hospital is still functional
•Print the current films for the ward or SOPD
•Ask the clinician come to radiology department for urgent cases
•Contact responsible parties as in Emergency Response Procedure
BCP triggered

Recovery Time for the DRP

During disaster recovery, timing was also important both for the staff and the manager. The recovery times of some critical subprocesses are listed as in the following table.

Backup Plan

Backup copies of important PACS system files, patient information, essential system information, and software should be made and tested regularly.

Security and Security-Awareness Training

Training (education concerning the vulnerabilities of the health information in an entity's possession and ways to ensure the protection of that information) includes all of the following implementation features.

i. Awareness training for all personnel, including management personnel (in security awareness, including, but not limited to, password mainte-

Table 2.

Step	Recovering Subprocess	Responsible Person	Process Location
1	Image distribution to clinicians	PACS team, contractor	Web servers
2	Image distribution through Cisco switches	PACS team, contractor	Cisco switches
3	Image online storage	PACS team, contractor	PACS servers, SAN
4	Image reporting	PACS team, contractor	Image viewers
5	Image prefetching from MOD jukebox	PACS team, contractor	MOD jukebox
6	Image prefetching from tape library	PACS team, contractor	Tape library
7	Image receiving	PACS team, radiographers, contractor	PACS servers
8	Image verification	PACS team, radiographers, contractor	PACS servers
9	Web-server load balancing	PACS team, contractor	Load-balancing switch
10	Image archiving to MOD jukebox	PACS team, contractor	Jukebox
11	Image archiving to tape library	PACS team, contractor	Tape library
12	Patient demographic-data retrieval	Radiographers, ITD	PACS broker
13	Remote maintenance	PACS team	RAS server and Cisco router

Table 3.

DRP Level Triggered	Scope	Recovery Time
1	Clinicians in a ward or the SOPD cannot view images while other parts of the hospital are still functional.	Half day for the recovering of subprocess no. 10
2	Clinicians in several wards or the SOPDs cannot view images while the PACS in the radiology department is still functional.	One day for the recovering of subprocess nos. 10 and 11
3	Neither the clinical department nor radiology can view images.	One week for the recovering of subprocess nos. 1 to 13

nance, incident reporting, and viruses and other forms of malicious software)

ii. Periodic security reminders (employees, agents, and contractors are made aware of security concerns on an ongoing basis)

iii. User education concerning virus protection (training relative to user awareness of the potential harm that can be caused by a virus, how to prevent the introduction of a virus to a computer system, and what to do if a virus is detected)

iv. User education in the importance of monitoring log-in success or failure and how to report discrepancies (training in the user's responsibility to ensure the security of healthcare information)

v. User education in password management (type of user training in the rules to be followed in creating and changing passwords and the need to keep them confidential)

Documentation and Documentation Control

Documentation and documentation control serve as a control on the document and data drafting, approval, distribution, amendment, obsolescence, and so forth to make sure all documents and data are secure and valid.

Standard and Legal Compliance

The purpose of standard and legal compliance (Hong Kong Personal Data Privacy Ordinance, 1995) was to avoid breaches of any criminal and civil law; statutory, regulatory, or contractual obligations; and any security requirements. Furthermore, the equipment compliance of the DICOM standard can improve the compatibility and upgradability of the system. Eventually, it can save costs and maintain data integrity.

Quality of PACS

In a filmless hospital, the PACS is a mission-critical system for lifesaving purposes. The quality of the PACS was an important issue. One method to measure the quality of a PACS was measuring the completeness of the system in terms of data confidentiality, integrity, and availability. A third-party audit such as the ISO17799/BS7799 certification audit could serve as written proof of the quality of a PACS.

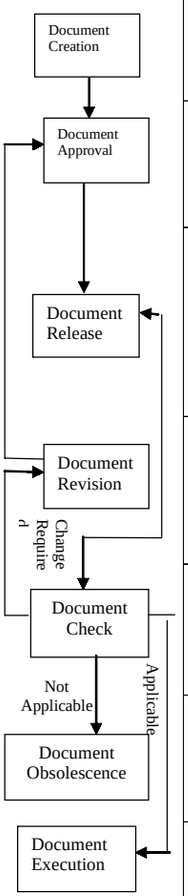
FUTURE TRENDS

Based on the experience in BS7799 implementation, the authors were of the view that more and more hospitals would consider similar healthcare applications of BS7799 to other safe-critical equipment and installations in Hong Kong.

CONCLUSION

ISO17799/BS7799 covers not only the confidentiality of the system, but also the integrity and availability of data. Practically, the latter is more important for the PACS. Furthermore, both standards can help to improve not only the security, but also the quality of a PACS because, to ensure the continuation of the certification, a security forum has to be established and needs to meet regularly to review and improve on existing processes.

Table 4.

Process Flow	Operation	Remark
	Manuals, procedures, and work instruction should be written by the PACS team. Records should be kept in the general office.	If documents/manuals cover different departments, we should consider liaisons between different departments' roles.
	Manuals should be approved by the chief of service (COS). Procedures and work instruction should be approved by the PACS manager. Records should be stored in the PACS room or general office.	Manual changes should be approved by the PACS manager.
	1. The distribution of manuals and procedures is controlled by the PACS manager. 2. The requirements from the customers and contracts related to information security of the PACS should be approved by the COS and released by the PACS manager.	Documents/manuals related to PACS should be signed by the PACS manager before distribution. The manual distributed should have a document number. Each personnel/department should update the document-control list regularly.
	Manuals and documents should be amended by the document owner/department. If other personnel/departments are involved in the change, they should seek the approval from the owner/responsible departments.	Note the change and where the change is (e.g., which paragraph) on the first page. The original document/manual should be chopped or destroyed.
	For general manuals from an outsourcing party (e.g., Afga) or other department, if they are applicable for PACS operation, they should be approved and adopted for PACS operation.	For this kind of manual, if it has not been revised for 1 year, it should be reviewed.
	Obsolete documents should be collected by the PACS manager. There should be one copy (soft copy or hard copy) kept by the PACS.	Each personnel/department should keep the previous updated version of the document for future review. The other obsolete copy should be destroyed.
	It should be guaranteed that the operator or other related PACS engineer should get the right document in the right version.	During operation, no document should be copied, duplicated, or distributed without appropriate approval.

REFERENCES

British Standards Institution. (2000). *Information technology: Code of practice for information security management* (BS ISO/IEC 17799: 2000 [BS 7799-1:2000]). UK: British Standards Institution.

British Standards Institution. (2002). *Information security management systems: Specification with guidance for use* (BS 7799-2 2002). UK: British Standards Institution.

Calder, A., & Watkins, S. (2003). *IT governance: A manager's guide to data security and BS 7799/ ISO 17799.* London: Kogan Page.

Dreyer, K. J., Mehta, A., & Thrall, J. H. (2001). *PACS: A guide to the digital revolution.* New York: Springer-Verlag.

Hong Kong Personal Data Privacy Ordinance. (1995). Hong Kong, China: Hong Kong Government.

Huang, H. K. (2004). *PACS and imaging informatics: Basic principles and applications.* Hoboken, NJ: Wiley-Liss.

The Institute of Physics and Engineering in Medicine. (2003). *Guidance notes on the recommendations for professional practice in health, informatics and computing.* UK and Institute of Physics and Engineering in Medicine.

Menasce, D. A., & Almeida, V. A. F. (2001). *Capacity planning for Web services: Metrics, models, and methods.* Upper Saddle River, NJ: Prentice Hall.

Peltier, T. R. (2001a). *Information security policies, procedures, and standards: Guidelines for effective information security management.* Boca Raton, FL: CRC Press.

Peltier, T. R. (2001b). *Information security risk analysis.* Boca Raton, FL: Auerbach Publishing.

A practical guide to IT security for everyone working in hospital authority. (2004). Hong Kong, China: Hong Kong Hospital Authority IT Department.

Security operations handbook. (2004). Hong Kong, China: Hong Kong Hospital Authority IT Department.

Toigo, J. W. (1996). *Disaster recovery planning for computer and communication resources.* John Wiley & Sons.

Toigo, J. W. (2003). *The holy grail of network storage management.* Prentice Hall.

KEY TERMS

Availability: Prevention of unauthorized withholding of information or resources.

Business-Continuity Planning: The objective of business-continuity planning is to counteract interruptions to business activities and critical business processes from the effects of major failures or disasters.

Confidentiality: Prevention of unauthorized disclosure of information.

Controls: These are the countermeasures for vulnerabilities.

Digital Imaging and Communications in Medicine (DICOM): Digital Imaging and Communications in Medicine is a medical image standard developed by the American College of Radiology and the National Electrical Manufacturers' Association.

Information-Security-Management System (ISMS): An information-security-management system is part of the overall management system, based on a business risk approach, to develop, implement, achieve, review, and maintain information security. The management system includes organizational structure, policies, the planning of activities, responsibilities, practices, procedures, processes, and resources.

Integrity: Prevention of unauthorized modification of information.

Picture-Archiving and -Communication System (PACS): A picture-archiving and -communication system is a system used for managing, storing, and retrieving medical image data.

Statement of Applicability: Statement of applicability describes the control objectives and controls that are relevant and applicable to the organization's ISMS scope based on the results and conclusions of the risk assessment and treatment process.

Threats: These are things that can go wrong or that can attack the system. Examples might include fire or fraud. Threats are ever present for every system.

Vulnerabilities: These make a system more prone to attack by a threat, or make an attack more likely to have some success or impact. For example, for fire, a vulnerability would be the presence of inflammable materials (e.g., paper).

Information Security Threats

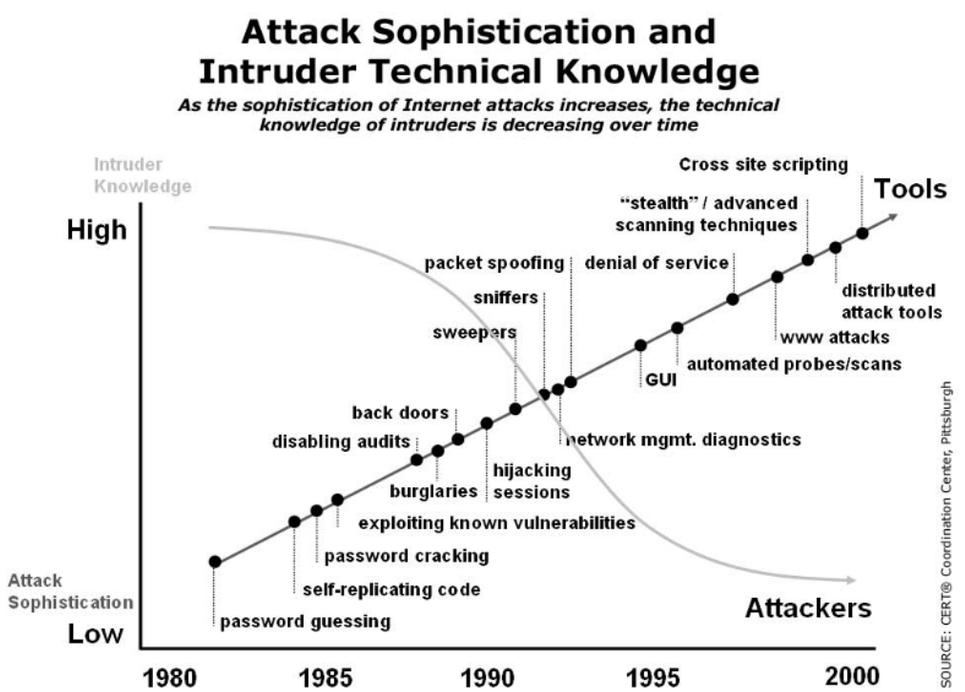
Rana Tassabehji
University of Bradford, UK

INFORMATION SECURITY: EVOLUTION TO PROMINENCE

Information security is an old concept where people, businesses, politicians, military leaders, and others have been trying to protect "sensitive" information from unauthorised or accidental loss, destruction, disclosure, modification, misuse, or access. Since antiquity, information security has been a decisive factor in a large number of military and other campaigns (Wolfram 2002)—one of the most notable being the breaking of the German Enigma code in the Second World War.

With the invention of computers, information has moved from a physical paper-based format to an electronic bit-based format. In the early days, mainframe infrastructures were based on a single sequential execution of programmes with no sharing of resources such as databases and where information could be relatively easily secured with a password and locked doors (Solms 1998). The development and widespread implementation of multi-processor personal computers and networks to store and transmit information, and the advent of the Internet, has moved us into an information age where the source of wealth creation is changing from atoms (physical goods), to bits (digital goods and services) (Negroponte 1995). Information is now a valuable asset and consequently, information security is increasingly under threat as vulnerabilities in systems are being exploited for economic and other gain. The CERT Co-ordination Center at Carnegie Mellon

Figure 1. Relationship between attack sophistication and knowledge required by attackers (Source: http://www.cert.org/archive/ppt/crime-legislation.ppt)



Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

University has charted the increase in sophistication of attacks as knowledge required decreases since technical attack tools are more readily available and indiscriminately accessible (Figure 1). Even novices can launch the most sophisticated attacks at the click of a mouse button (Anthes 2003).

MEASURING INFORMATION SECURITY THREATS

It is impossible to get accurate figures for the number and cost of information security breaches, mainly because organisations are either not aware that the breach has occurred, or are reluctant to publicise it, for fear of ruining their reputation or destroying the trust of their stakeholders. However, in one instance the impact of malicious software in the form of worm/virus attacks on the Internet was estimated to have caused $32.8 billion in economic damages for August 2003 (Berghel 2003).

The types of information security threats come from a number of sources, which can be broadly divided into two main categories the technical and non-technical which will be examined in more detail in the next section.

TECHNICAL INFORMATION SECURITY THREATS

Information is increasingly transmitted and stored on interconnected and networked infrastructures, so the threats to information security can come from a variety of technical sources, including:

- **Intrusion attacks:** where hackers or unauthorised intruders gain access to stored information to either steal or vandalise it (such as defacing a Web site).
- **Probing or scanning:** where an automated tool is used to find and exploit vulnerabilities to gain access to an information system as a prelude to theft or modification of information. Increasingly home users are unknowing victims of scans that detect unprotected ports allowing attackers to gain access to their information or take control of their computer to launch other types of attack.
- **Automated eavesdropping:** uses sniffer programmes that monitor and analyse information in transit. They capture information such as usernames, passwords, or other text being transmitted over a network, which might not always be encrypted, for instance e-mail.
- **Automated password attacks:** the most common and successful kind of threat to information security. They exploit people's poor password practices (see Table 1) and their tendency to rely on passwords that are easy to remember and have some personal relevance. Once attackers have a user's password, they can legitimately access all their privileges and information. Some techniques used include:
  - **Brute-force attacks:** where programmed scripts run through every single combination of characters until the password is found. This takes time and is the slowest method since for an eight-character lower case alphabet there are 200 billion combinations, but powerful processors can reduce the time considerably. It is most effective for short and simple passwords.
  - **Dictionary attacks:** where programmed scripts run through every word in one or more dictionaries that include different languages, common names, and terms from popular culture such as films, music, or sport until the password is found.
  - **Password cracking:** a more advanced method where attackers launch a brute-force dictionary attack to find out encrypted passwords. Common words (found in dictionaries) are encrypted and compared to stored encrypted passwords until a match is found. The success of this attack depends upon the completeness of the dictionary of encrypted passwords and the processor power of the machines being used.
- **Spoofing:** where a person or machine impersonates another to gain access to a resource, making it easy for an attacker to modify original information or change its destination. This technique is effective in disguising an attacker's identity, preventing victims from identifying the culprits who breach their systems. A more recent trend is *"phishing",* where the mass distribution of "spoofed" e-mail messages with

Table 1. Examples of poor password practices

Common Password Practices
Writing down a password and placing it on or near the computer.
Using a word found in a dictionary followed by a couple of numbers
Using names of people, places, pets, common items or birth dates
Sharing password – managers with secretaries and work colleagues are particularly guilty of this practice.
Using the same password for more than one account, and for an extended period of time
Using the default password provided by the vendor.

return addresses, links, and branding appear to come from legitimate banks, insurance companies, retailers but are fraudulent (Anon, 2004; Barrett, 2004). The e-mails request personal information, credit card/pin, or account numbers. The majority of phishing attacks use a link to a fraudulent Web site or ask the recipient to download a file that contains some form of malware. Ebay, PayPal and a number of international banks have consistently appeared in the list of most targeted companies compiled by the Anti-Phishing Working Group in 2004 (www.antiphising.org).

- **Denial of Service (DoS) attacks:** exploit weaknesses in the design of information systems and come in different forms. They mainly involve the sending of an excessively large number of data packets to a destination that it is unable to handle the requests, which ultimately brings the system down. Some can also contain codes designed to trigger specific actions, for example, damage files, change data, or disclose confidential information. This causes maximum disruption and cost by depriving legitimate users of normal network services.
- **Malware:** also known as malicious software, is the most common and high profile type of attack specifically designed to cause harm in the form of viruses, Trojan horses, worms, Visual Basic, and Java scripts that hide in some Web pages and execute pre-programmed commands when activated.

A virus is a manmade program code often designed to automatically spread to other computer users. The payload, or consequences, of each virus depends on the code written within it. Some viruses are harmless, for example the William Shakespeare virus activated on April 25 displays the message "Happy Birthday, William!" Others can be very harmful, erasing, or corrupting data, re-formatting hard drives, e-mailing private or sensitive information to address book listings, or installing spyware that e-mails passwords or other confidential information to unauthorised recipients. The warnings of Professor Cohen are still relevant today:

Viral attacks appear to be easy to develop in a very short time, can be designed to leave few if any traces ... and require only minimal expertise to implement. Their potential threat is severe, and they can spread very quickly. (Cohen 1984)

Around 60,000 viruses have already been identified and 400 new ones are being created each month (Trend Micro, 2004). Whereas viruses previously were spread by floppy disks, and attacked one file at a time; in the digital age, viruses utilise systems, networks (including the Internet) and e-mail programmes to replicate themselves rapidly and exponentially. They now circumvent the advice to beware e-mails that come from unknown users, since the majority of viruses use seemingly legitimate and trusted recipients sourced from the user's own address book.

NON-TECHNICAL TYPES OF INFORMATION SECURITY THREATS

In the past, much information security research and attention focussed largely on technical issues. However, in recent years, it has become widely acknowledged that human factors play a part in many security failures (Weirich & Sasse, 2002; Whitman, 2004). While technical threats are usually more high pro-

file and given much media and financial attention, non-technical human and physical threats are sometimes more effective and damaging to information security. Non-technical threats include:

- **"Acts of God":** such as fire, flood, and explosion—both paper and bit-based information could be permanently destroyed and impossible to recover or recreate.
- **Physical infrastructure attacks:** such as theft or damage of hardware, software, or other devices on or over which information is stored or transmitted. This could lead to permanent loss or unauthorised access to critical information.
- **Acts of human error or failure:** where operators make genuine mistakes or fail to follow policy (Loch, Carr, et al., 1992).
- **Social engineering:** uses human interaction to break security procedures. This might involve gaining the confidence of employees with access to secure information; tricking them into thinking there is a legitimate request to access secures information; physical observation; and eavesdropping on people at work. Social engineering preys on the fact that people are unable to keep up with the rapid advance of technology and little awareness of the value of information to which they have access. Kevin Mitnick (Mitnick & Simon, 2003), one of the most high-profile "hackers", underlined the importance of social engineering in obtaining access to systems:

When I would try to get into these systems, the first line of attack would be what I call a social engineering attack, which really means trying to manipulate somebody over the phone through deception. I was so successful in that line of attack that I rarely had to go towards a technical attack. The human side of computer security is easily exploited and constantly overlooked. Companies spend millions of dollars on firewalls, encryption and secure access devices, and it's money wasted, because none of these measures address the weakest link in the security chain.

US Senate Testimony (Mitnick 2000)

Bruce Schneier, one of the world's leading security experts, similarly underlines the importance of social engineering: "amateurs hack systems, professionals hack people" (Christopher 2003).

A DISCUSSION OF INFORMATION SECURITY THREATS

None of the threats mentioned are mutually exclusive and could occur in any combination. All threaten the information and the systems that contain and use them.

Although there can be no agreement on the actual figures and percentages, empirical evidence from a number of security surveys over the past years (CompTIA, 2003; CompTIA, 2004; Pricewater houseCoopers,2002; PricewaterhouseCoopers, 2004; Richardson, 2003) shows similar trends and patterns of security breaches. Information security breaches are increasing year on year. The most common type of attack is from viruses and malware, followed by hacking or unauthorised access to networks resulting in vandalism of Web sites and theft of equipment (mainly laptops). Denial-of-service attacks are less frequent relative to viruses, with financial fraud and theft of information being the lowest kind of security breach experienced. However, it should be noted that the latter two breaches would be hard to detect in the short term and the impact of the previous attacks would have an indirect effect on the information stored. It is commonly believed that information security is most at risk from insiders, followed by ex-employees, hackers, and terrorists to a lesser extent (PricewaterhouseCoopers,2002; Pricewater houseCoopers, 2004).

Schultz (2002) argues that there are many myths and misconceptions about insider attacks and develops a framework for predicting and detecting them in order to prevent them. Although this framework has not yet been validated by empirical evidence, the metrics identified are drawn from a range of studies in information security by a number of academics. Some of the measures identified are personality traits; verbal behaviour; consistent computer usage patterns; deliberate markers; meaningful errors; and preparatory behaviour (Schultz, 2002). In academic terms, the field of information security is still young and this is one area in which more research can be conducted.

FUTURE TRENDS

It is always difficult to predict the future, but the past and present allows us some insight into trends for the future. Over the last few years, information security has changed and matured, moving out of the shadow of government, the military and academia into a fully fledged commercial field of its own (Mixter, 2002) as the commercial importance and economic value of information has multiplied.

Information is reliant on the systems that manage and process it. The future trend for information systems technology is more intelligent information processing (in the form of artificial intelligent bots and agents) and the increased integration and interoperability between systems, languages, and infrastructures. This means a growing reliance on information in society and economy and a subsequent rise in importance of information security.

In the short term, nobody predicts that there will be a termination of information security threats. There will be an escalation of blended combined threats with more destructive payloads—for instance, the development of malware that disables anti-virus software, firewalls, and anti-Trojan horse monitoring programmes (Levenhagen, 2004). Although the measures being taken to protect information will continue to be a cocktail of procedures in the short term, there are two views of how the threat to information security will develop in the longer term.

On the one hand, there are those that feel information security will improve incrementally as vulnerabilities are tackled by researchers and businesses. A study into the history of worms (Kienzle & Elder, 2003) identified the process of creating worms as evolutionary and that best security practices do work against this threat. Mixter (2002) and others (Garfinkel, 2004; Kienzle & Elder, 2003) know there is still much work to be done, but identify the need for information security to define clear rules and guidelines for software developers while also improving user intelligence and control. The main areas for potential research not yet fully explored, are the development of new approaches to information security education and policies, trust and authentication infrastructures, intelligence, and evaluation to quantify risk in information systems. None of which are easy.

On the other hand, there is the *"digital Pearl Harbour"* view, which posits that information security will only improve as a result of an event of catastrophic and profoundly disturbing proportions (Berinato, 2003; Schultz, 2003) that will lead to the mobilisation of governments, business, and people. The consequences of the *"digital Pearl Harbour"* would lead to a cycle of *recrimination* where the first response will be litigation of those that are liable. *Regulation* would follow with the rapid introduction of legislation to counter or prevent the catastrophe and the introduction of standards for software development. These would include configuration of software; reporting vulnerabilities; common procedures for virus or other attacks. Finally, *reformation* would change attitudes to information security and there would be a cultural shift for a better and more pro-active approach with zero tolerance for software that threatens information and system security (Berinato, 2003). Alternatively, the reaction to the *"digital Pearl Harbour"* would be to remove the integration between systems enforcing security restrictions that do not allow information sharing or transmission. Some (Garfinkel, 2004) predict that if the issue of information security is not resolved the use of new technology for sharing information (such as e-mail) will become a mere footnote of communications history, similar to the CB radio.

CONCLUSION

Information is now the lifeblood of organisations and businesses—some even argue the economy. In order to grow and thrive, information must be secured. The three most common features of information security that are threatened by both technical and non-technical means are ensuring:

- **Confidentiality:** That information is accessible only to those authorised to access it.
- **Integrity:** That information is unchanged and in its original format whether it is stored or transmitted, and being able to detect whether information has been tampered with, forged or altered in any way (whether accidentally or intentionally).

- **Authentication:** That the source of the information (whether individuals, hardware, or software) can be authenticated as being who they claim to be.

But there must also be *accountability* and *authorisation*, where security protocols and procedures are clearly defined and can be traced and audited. The information security threats described, are just a sample of the kinds of attack that can occur. They all underline the fact that information security in the digital and interconnected age is heavily reliant on technology. However, the technology being developed to share and transmit information has not been able to keep up with the types of threats that have emerged. This lack of progress is dependent on a combination of different factors.

- Security has not been a design consideration but an afterthought, as "patches" are bolted on after vulnerabilities have been exploited.
- Legislation for those that breach security and development of common technical standards, has still to be developed.
- Education and awareness-raising for users to improve "computing" and information security practices, has been lagging behind the rapid and widespread implementation and use of the new digital infrastructure.

Information security is not solely a technology issue. The kinds of vulnerabilities that exist in people's working practices, hardware, software, and the infrastructure of the Internet and other systems as a whole, are many and so information security is the responsibility of all the stakeholders and any measures to combat information security threats should be a combination of the technical and non-technical.

REFERENCES

Anon (2004). Internet "Phishing" scams soared in April. *Wall Street Journal (Eastern Edition)*. New York.

Anthes, G. (2003). Digital Defense. *Computerworld, 37*(51), 32.

Barrett, J. (2004). When crooks go Phishing. *Newsweek, 143*, 66.

Berghel, H. (2003). Malware month. *Communications of the ACM, 46*(12), 15.

Berinato, S. (2003). The future of security. *CIO, 17*(6), 1.

Christopher, A. (2003). The Human Firewall. *CIO*. Retrieved 28/10/2003 from *http://www.CIO.co.nz*

Cohen, F. (1984). Experiments with Computer Viruses, *http://www.all.net/books/virus/part5.html, accessed 24/3/2004*

CompTIA (2003). Committing to Security: A CompTIA Analysis of IT Security and the Workforce. ComputingTechnologyIndustryAssociation, *http://www.comptia.org/research/files/summaries/SecuritySummary031703.pdf, accessed 24/3/2004*

CompTIA (2004). Computer Viruses, Worms Pose Biggest Security Headache for IT Departments. ComputingTechnologyIndustryAssociationWebPoll, *http://www.comptia.org/pressroom/get_news_item.asp?id=364, accessed 24/3/2004*

Garfinkel, S. (2004). Unlocking our future: A look at the challenges ahead for computer security. Machine Shop - technologies, tools and tactics. E. Cummings, CSO Magazine, *http://www.csoonline.com/read/020104/shop.html*

Kienzle, D.M. & Elder, M.C. (2003). Internet WORMS: past, present and future; Recent worms: a survey and trends. *Proceedings of the 2003 ACM Workshop on Rapid Malcode*, Washington.

Levenhagen, R. (2004). Trends, codes, and virus attacks: 2003 year in review. Network security, *2004*(1), 13-15.

Loch, K.D., Carr, H.H., et al. (1992). Threats to information systems: Today's reality, yesterday's understanding. *MIS Quarterly, 16*(2), 173-86.

Mitnick, K. (2000). Senate Governmental Affairs Committee, *http://www.kevinmitnick.com/news-030300-senatetest.html, accessed 10/3/2004*

Mitnick, K. & Simon, W.B. (2003). *The art of deception: Controlling the human element of security.* John Wiley & Sons.

Mixter (2002). (D)evolution of Information Security and Future Trends, *http://mixter.warrior2k.com/is-evol.html, accessed 20/3/2004*

Negroponte, N. (1995). *Being digital*. Alfred A. Knopf.

PricewaterhouseCoopers (2002). Department of Trade and Industry: Information Security Breaches Survey, *http://www.security-survey.gov.uk/, accessed 24/3/04*

PricewaterhouseCoopers (2004). Department of trade and industry: Information security breaches survey, *http://www.security-survey.gov.uk/, accessed 24/3/2004*

Richardson, R. (2003). CS & FBI computer crime and security survey. *Computer Security Institute, 21*.

Schultz, E.E. (2002). A framework for understanding and predicting insider attacks. *Computers & Security, 21*(6), 526-531.

Schultz., E.E. (2003). Internet security: What's in the future? *Computers & Security, 22*(2), 78-79.

Solms, O.V. (1998). Information Security Management (1): Why information security is so important. *Information Management & Computer Security, 6*(4), 174-177.

Weirich, D. & Sasse, M.A. (2002). Pretty good persuasion: A first step towards effective password security in the real world. *Association of Computing Machinery NSPW'01*, New Mexico.

Whitman, M.E. (2004). In defense of the realm: Understanding the threats to information security. *International Journal of Information Management, 24*(1), 43-57.

Wolfram, S. (2002). A New Kind of Science. *Wolfram Media, 1085*.

KEY TERMS

Biometrics: The science of measuring, analysing, and matching human biological data such as fingerprints, irises, and voice/facial patterns. In information system security, these measures are increasingly being introduced for authentication purposes and will play a critical role in the future of digital security.

Crackers: Coined in the 1980s by hackers wanting to distinguish themselves from someone who intentionally breaches computer security for profit, malice, or because the challenge is there. Some breaking-and-entering has been done ostensibly to point out weaknesses in a security system.

Cryptography: Protecting information by transforming it into an unreadable format using a number of different mathematical algorithms or techniques.

Firewall: A combination of hardware and software that prevents unauthorised access to network resources—including information and applications.

Hackers: A slang term for a computer enthusiast or clever programmer, more commonly used to describe individuals who gain unauthorised access to computer systems for the purpose of stealing or corrupting information or data. Hackers see themselves as the "white hats" or the good guys who breach security for the greater good. The media at large makes no distinction between hackers and crackers.

Phishing: Scams use e-mail and Web sites designed to look like those of legitimate companies, primarily banks, to trick consumers into divulging personal information, such as financial account numbers, that can be used to perpetrate identity-theft fraud (http://www.antiphishing.org/)

Port: An interface for physically connecting to some other device such as monitors, keyboards, and network connections.

Trojan Horse: A program in which malicious or harmful code is disguised as a benign application. Unlike viruses, Trojan horses do not replicate themselves but can be as destructive.

Worm: A program or algorithm that resides in active memory and replicates itself over a computer network, usually performing malicious actions, such as using up the computer's resources and shutting down systems. Worms are automatic and are only noticed when their uncontrolled replication has used so much of a system's resources that it slows or halts other tasks.

Information Systems Strategic Alignment in Small Firms

Paul B. Cragg
University of Canterbury, New Zealand

Nelly Todorova
University of Canterbury, New Zealand

INTRODUCTION

The concept of "alignment" or "fit" expresses an idea that the object of design—for example, an organization's structure or its information systems (IS)—must match its context to be effective (Iivari, 1992). More recently, Luftman (2004) has taken this argument one step further and argued that a lack of alignment within an organization will limit the effectiveness of the organization's business strategies.

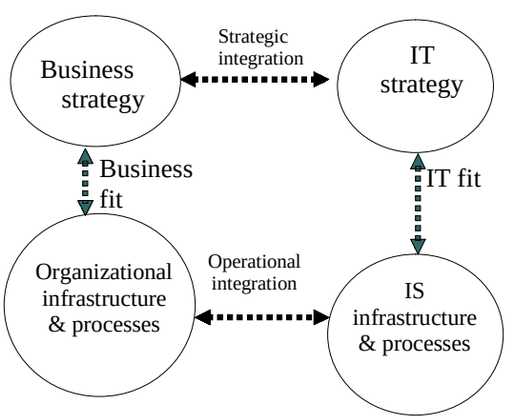
The concept of alignment has become particularly important in the field of IS, as Luftman (2004) and others have argued that firms need to align their IS strategies with the other strategies of the business. For example, if a firm's business strategy is to be a "cost leader" in its industry, then its IS strategies should support and enable "cost leadership;" for example, through effective supply chain management.

Much of the research on IT alignment builds on the work of Henderson and Venkatraman (1989), who identified four types of alignment within organizations. They developed a strategic alignment model that defined the range of strategic choices facing managers and how they interrelate. Their model is summarized in Figure 1, with four domains of strategic choice: business strategy, IT strategy, organizational infrastructure and IT infrastructure. They argue that alignment requires organizations to manage the fit between strategy and structure, as well as the fit between the business and IT. They named the four aspects of alignment as:

- **Strategic integration** – the alignment between business and IT strategies
- **Operational integration** – the alignment between business infrastructure and IT infrastructure
- **Business fit** – the alignment between business strategy and business infrastructure
- **IT fit** – the alignment between IT strategy and IT infrastructure

Typically, different researchers have focused on parts of the Henderson & Venkatraman (1993) model. For example, Chan, Huff, Barclay and Copeland (1997) focused on the link between business strategy and IT strategy, while Raymond et al. (1995) focused on the link between organizational structure and IT structure. Most of the recent research has focused on Henderson & Venkatraman's (1989) "strategic integration"; that is, alignment at the strategy level. This type of alignment is now typically referred to as "strategic alignment." This article focuses on strategic alignment, partly because there has been significant research in recent years that has focused on strategic alignment, but also because recent research indicates that alignment at the strategic level is important for all organizations that use IT.

Figure 1. The Henderson and Venkatraman strategic alignment model (1993)



Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

STRATEGIC ALIGNMENT

Despite the wide recognition of the importance of IT alignment, studies have indicated that firms struggle to achieve alignment (Chan et al., 1997; Luftman 2004). For example, Luftman (2004) places most large firms that he has studied at an IT alignment maturity level of 2, on his scale from 1 to 5, where 1 is least mature/not aligned and 5 indicates mature/fully aligned. As a result, some researchers have examined factors that influence IT alignment in an attempt to understand how firms can best achieve alignment. In particular, Reich and Benbasat (2000) concentrated on the antecedents that influence alignment. In their study, they used the duality of strategy creation: an intellectual and a social dimension. The intellectual dimension refers to methods and techniques, while the social dimension refers to people involved and their role. Reich and Benbasat defined the social dimension of IT alignment as, "the state in which business and IT executives within an organizational unit understand and are committed to the business and IT mission and objectives." Reich and Benbasat (2000) identified five major factors that influenced the social dimension of IT alignment: shared domain knowledge between business and IT executives, IT implementation success, communication between business and IT executives, connections between business and IT planning processes, and strategic business plans.

Luftman (2004) is another who has focused on enablers of alignment in firms, resulting in the following six enablers of IT alignment: communications between IT and the business, IT/business value measurements, IT governance, IT partnerships, IT scope and architecture and IT skills. Luftman (2004) outlines the content of each enabler. For example, "communication" includes six aspects, including communication by IS staff with the rest of the business and communication by the rest of the business with IS. He argues that all six enablers contribute to "alignment maturity," and he encourages firms to evaluate all six enablers, then create project plans to improve the organization's level of alignment.

The studies by Reich and Benbasat (2000) and Luftman (2004) show that alignment is influenced by a broad range of factors and that we have yet to reach a consensus on these factors. Importantly, both IT and non-IT managers and staff can influence alignment. They all make important contributions, so they must work as a partnership.

Although IT alignment has been discussed by many, there have been relatively few attempts to measure IT alignment. Chan et al. (1997) conducted one of the most comprehensive attempts to quantify alignment and its effect on organizational performance. Chan et al. (1997) developed four survey instruments to measure each of the following constructs: business strategy, IS strategy, IS effectiveness and business performance. Venkatraman's (1989b) STROBE instrument was adapted for the business strategy instrument. A similar instrument was developed by Chan to assess IS strategy. As both instruments used the same eight dimensions of strategy, the two instruments were used to compute strategic fit. Chan found that alignment was a better predictor of performance than the individual measures of strategy, and thereby demonstrated a positive relationship between strategic alignment and business performance.

There is also some debate about how data should be analyzed when attempting to measure alignment. Matching and moderation are two of the many ways of measuring alignment (Hofacker, 1992). The matching perspective is commonly based on the difference between two measures. For example, if "cost reduction" was rated by a firm as having an importance of 10, and the IT support for "cost reduction" had a rating of 2, then the matching approach would use the absolute difference of 8 (i.e., 10 – 2), as an indication of the alignment of IT with the "cost reduction" strategy. Using the matching approach, alignment is thus the level of similarity between the measures.

Another common perspective is "moderation," which assumes that alignment reflects synergy; for example, between IS and business strategy. Alignment is thus calculated as the interaction between the two measures. For example, if "cost reduction" was rated by a firm as having an importance of 10, and the IT support for "cost reduction" had a rating of 2, then the moderation approach would give this a score of 20 (i.e., 10 * 2), as an indication of the alignment of IT with the "cost reduction" strategy. The moderation perspective gives greater weight to, for example, a firm's most important business strategies.

Chan et al.'s (1997) results supported the moderation approach. Bergeron et al. (2001) explored six perspectives of alignment and found support for

both the matching and moderation approaches in part of their model, but concluded that more research was needed on the different perspectives.Cragg, King and Hussin (2002) found support for the moderation approach, as well as evidence that the matching approach could provide misleading results. Sometimes the matching approach could indicate high alignment when other indicators suggested that alignment was not high.

IT ALIGNMENT IN SMALL FIRMS

Most of the research to date on IT alignment is based on the experiences of large firms. IT alignment in small firms has yet to receive much attention, although there is evidence that IT alignment exists in small firms. For example, Levy, Powell and Yetton (1998) identified "innovation" firms, where "IS are an integral and tightly woven part of the business strategy" (p. 6). They also provided evidence of a lack of IT alignment in their "efficiency" firms, where "there is no recognition of the role of information in supporting the achievement of business strategy" (p. 5). These results have been supported since by Cragg et al. (2002), who reported a high degree of alignment between business strategy and IT strategy for a significant proportion of the small manufacturers that they examined. Furthermore, the group of small firms with high IT alignment had achieved better organizational performance than firms with low IT alignment.

Some researchers have examined how small firms can align their IT strategy with their business strategy. In particular, Blili and Raymond (1993) argued that small firms must adopt some kind of framework for planning IT if they wish to create IT-based strategic advantage. Subsequently, Levy and Powell (2000) proposed an approach to IS strategy (ISS) development aimed specifically at small firms, to help them align their IT and business strategies. Their approach includes both business and IS planning and thus should encourage IT alignment. They have yet to report an evaluation of the effectiveness of their ISS development approach, including its impact on alignment. Furthermore, although Hussin, King and Cragg (2002) found the CEO's software knowledge to influence alignment, their personal involvement in IT planning and their IT use seemed to have relatively little influence on IT alignment. Their results indicated that the key influences on IT alignment are IT maturity and technical IT sophistication.

As well as enablers to IT alignment, other studies indicate factors that could inhibit IT alignment in small firms. Many studies have indicated that small firms do not have the resources to use IT in a strategic way. For example, managers in small firms are few in number, and have limited time and IT expertise, which limits their ability to devise IT strategy (Mehrtens, Cragg & Mills, 2001). Also, Hagmann and McCahon (1993) claim that small firms tend not to develop IS strategies. Consequently, this results in a lack of appropriate policies towards IT assessment and adoption, which reduces the likelihood of IT alignment. Furthermore, Palvia, Means and Jackson (1994) argued that the computing environment in very small firms (with 50 or less employees) was fundamentally different from medium-size firms, where there was often a formal IS department and a community of end users.

As with studies of large firms, there is no agreed way to measure IT alignment in small firms. The only significant attempt to date was reported by Hussin et al. (2002), who focused on the support by IT for nine aspects of business strategy. These items reflected: pricing, product quality, service quality, product differentiation, product diversification, new product, new market, intensive marketing and production efficiency. They used these items to identify that many firms had high IT alignment. However, their analysis found support for only seven of the items. Thus, their instrument requires further validation.

Ravarini, Tagliavini and Buonanno (2002) have also examined IT alignment in small firms as part of their instrument devised to provide an "IS check-up" for a small firm; that is, an instrument aimed at assessing the health of a small firm's IT. Their methodology includes an assessment of strategic alignment based primarily on the IT fit for each part of a small firm; for example, the sales area, accounting, logistics and so forth. Their assessment would evaluate the actual IT support in each area and the potential for IT for the area. Thus, their IT alignment is more at the operational level than Hussin et al.'s (2002) strategic integration, based on the Henderson and Venkatraman (1993) model. Their exploratory application of the model in small firms indicates that some units within small firms are well supported by

IT, but there are plenty of opportunities for IT to play a greater role within small firms.

FUTURE TRENDS

As yet, we still know too little about IT alignment in small firms to offer much advice to managers of small firms. One of the most important research opportunities is to design a valid and reliable way of measuring IT alignment in small firms. A valid instrument will enable the study of many other aspects of IT alignment and provide a tool for managers of small firms. The instrument by Hussin et al. (2002) could be developed further through rigorous validation, as they reported mixed results for their nine strategy items so they used seven in their analysis. It may also be possible to adapt the instruments used by Ravarini et al. (2002) and/or Chan et al. (1997).

Other ways of measuring fit could also be developed, based on the Henderson and Venkatraman (1993) model. For example, they indicated four types of alignment. Hussin et al. (2002) focused solely on one of these; that is, the alignment between business and IT strategy. Further research could focus on other aspects of alignment in small firms. Also, even when studying alignment at the strategic level, it may be beneficial for a study to focus solely on a firm's dominant business strategy. This focused approach could provide the opportunity to examine IT alignment with a particular business strategy; for example, service quality, to understand how service quality is best supported by IT. Some strategies may be easier to support than others. Also, some firms may be targeting IT at specific strategies.

Many small firms have achieved a high degree of alignment between their business strategy and IT (Cragg et al., 2002). However, we know very little about how this alignment was achieved. It may or may not have been planned using systematic frameworks, as argued by Blili and Raymond (1993) and Levy and Powell (2000). It seems more likely that the IT planning was informal (Lefebvre & Lefebvre, 1988). Further research could examine how small firms achieve IT alignment, and whether planning methodologies can be used to increase IT alignment in small firms.

Prior studies in large firms show that alignment is influenced by a broad range of factors (Luftman et al., 2004; Reich & Benbasat, 2000). Their findings could be examined in the context of small firms with the aim of identifying enablers of IT alignment that apply to small firms. For example, both IT and non-IT managers influence alignment in large firms. However, most small firms do not have IT managers or an IT department. Thus, small firm alignment in small firms requires further study, as it seems likely that not all of the factors identified by Luftman et al. (2004) and Reich and Benbasat (2000) are applicable to small firms. Less-formal aspects may be significant within small firms. For example, the multiple responsibilities taken on by some managers within small firms could mean that many managers are involved in strategy development. This would make it easy to share ideas about opportunities for IT, and thus foster connections between business and IT planning processes.

Cragg et al. (2002) used ANOVA to identify a positive association between IT alignment and small firm organizational performance. They used four measures of performance, including profit and sales. Performance was consistently higher in the group of firms that were most highly IT aligned. While they did not claim a causal link, the results were consistent with studies of larger organizations (Chan et al., 1997; Burn, 1996). The result also indicates that IT alignment could be a key to understanding the relationship between IT and firm performance. This is an area worthy of more research; that is, to better understand any relationship between alignment and outcomes like IT impact and firm performance. If alignment influences performance, what are the causal links? For example, does a lack of alignment lead to resources being wasted on non-productive activities; for example, more time spent seeking data? If we can understand the relationship better, then this is likely to indicate the ways that IT alignment could be improved to assist small firms.

The results of high alignment by some small firms imply that some small firms manage IT differently (Cragg et al., 2002). It seems possible that these varying levels of IT alignment are a reflection of " orientations"; that is, ways that managers and employees within firms view and treat IT, based on Venkatraman's (1989b) "strategic orientations" of firms. The generic IS linking strategies proposed by Parsons (1983) may provide a good starting point for identifying "IT orientations." Some of Parsons' strategies may apply to small firms, particularly centrally planned, scarce resource and necessary evil. Also,

Berry (1998) proposed a strategic planning typology for small firms. Furthermore, Joyce, Seaman and Woods (1998) identified "strategic planning styles" linked to process and product innovation in small firms. Importantly, these or other "IT orientations" may reflect IS cultures that have strong influences on IT alignment in small firms. As yet, "IT orientation" has not been researched in small firms.

CONCLUSION

The topic of IT alignment has received some attention in recent years, because studies have indicated that alignment has the potential to help improve our understanding of links between the deployment of IT and organizational effectiveness. To develop a better understanding of the concept of IT alignment and how it can be achieved, previous studies have investigated enabling factors, measures to quantify IT alignment and the development of processes to achieving alignment. Although theoretical frameworks have been proposed for IT alignment, relatively few have been discussed in relation to small firms. Previous research shows that frameworks developed for large firms cannot be applied directly to small firms. This paper suggests that some of the enabling factors may not be applicable to small firms, where managers have multiple roles and the planning process is more informal. Conversely, there could be additional factors affecting IT alignment, as previous research shows that many small firms lack the time and IT expertise for strategic application of IT, and do not develop IT strategies. Another major research opportunity is the development of an instrument that can be shown to measure IT alignment in small firms. Such an instrument would enable studies that examine the processes by which some small firms achieve IT alignment, as well as relationships between IT alignment and dependent variables like IT impact and organizational performance.

REFERENCES

Bergeron, F., Raymond, L. & Rivard, S. (2001, April). Fit in strategic information technology management: An empirical comparison of perspectives. *Omega, 2*(29), 124-142.

Berry, M. (1998). Strategic planning in small high tech companies. *Long Range Planning,* (3), 455-466.

Blili, S., & Raymond, L. (1993). Information technology – Threats and opportunities for small and medium-sized enterprises. *International Journal of Information Management, 13*, 439-448.

Burn, J.M. (1996). IS innovation and organizational alignment – a professional juggling act. *Journal of Information Technology, 11*, 3-12.

Chan, Y.E., Huff, S.L., Barclay, D.W., & Copeland, D.G. (1997). Business strategic orientation, information systems strategic orientation and strategic alignment. *Information Systems Research,* (2), 125-150.

Cragg, P., King, M., & Hussin, H. (2002). IT alignment and firm performance in small manufacturing firms. *Journal of Strategic Information Systems,* (2), June, 109-132.

Hagmann, C., & McCahon, C. (1993). Strategic information systems and competitiveness. *Information & Management, 25*, 183-192.

Henderson, J.C., & Venkatraman, N. (1993) Strategic alignment: A model for organizational transformation through information technology. *IBM System Journal,* (1), 4-16.

Hofacker, C.F. (1992). Alternative methods for measuring organization fit: Technology, structure and performance. *MIS Quarterly,* (1), March, 45-57.

Hussin, H., King, M., & Cragg, P. (2002). IT alignment in small firms. *European Journal of Information Systems,* (2), June, 108-127.

Iivari, J. (1992). The organizational fit of information systems. *Journal of Information Systems, 2,* 3-29.

Joyce, P., Seaman, C., & Woods, A. (1996). The strategic management styles of small businesses. In R. Blackburn & P. Jennings (Eds.), *Small firms: Contributions to economic regeneration* (pp. 49-58). London: Paul Chapman.

Lefebvre, E., & Lefebvre, L. (1992). Firm innovativeness and CEO characteristics in small manufacturing firms. *Journal of Engineering and Technology Management, 9,* 243-277.

Lefebvre, L.A., & Lefebvre, E. (1988). Computerization of small firms: A study of the perceptions and expectations of managers. *Journal of Small Business and Entrepreneurship,* (5), 48-58.

Levy, M., & Powell, P. (2000). Information systems strategy for small and medium sized enterprises: an organizational perspective. *Journal of Stragetic Information Systems,* (1), March, 63-84.

Levy, M., Powell, P., & Yetton, P. (1998). SMEs and the gains from IS: from cost reduction to value added. *Proceedings of the IFIP*, Helsinki, August 2-6.

Luftman, J.N. (2004). *Managing the information technology resource.* Upper Saddle River, NJ: Pearson Education.

Mehrtens, J., Cragg, P.B., & Mills, A.J. (2001). A model of internet adoption by SMEs. *Information & Management,* (3), Dec, 165-176.

Palvia, P., Means, D.B., & Jackson, W.M. (1994). Determinants of computing in very small businesses. *Information & Management, 27*, 161-174.

Parsons, G.L. (1983, Fall). Information technology: A new competitive weapon. *Sloan Management Review, 25*(1), 3-14.

Porter, M.E. (1980). *Competitive strategy – Techniques for analysing industries and competitors.* New York: Free Press.

Ravarini, A., Tagliavini, M., & Buonanno, G. (2002). Information system check-up as a leverage for SME development. In S. Burgess (Ed.), *Managing IT in small business* (pp. 63-82). Hershey, PA: Idea Group Publishing.

Raymond, L., Pare, G., & Bergeron, F. (1995). Matching information technology and organizational structure: An empirical study with implications for performance. *European Journal of Information Systems, 4*, 3-16.

Reich, B.H., & Benbasat, I. (2000). Factors that influence the social dimension of alignment between business and information technology objectives. *MIS Quarterly,* (1), March, 81-111.

Venkatraman, N. (1989a). The concept of fit in strategy research: Toward verbal and statistical correspondence. *Academy of Management Review, 14*(3), 423-444.

Venkatraman, N. (1989b). Strategic orientation of business enterprises – The construct, dimensionality, and measurement. *Management Science, 35*(8), 942-962.

KEY TERMS

Business Performance: Reflects an organization's overall results and is often measured using a number of financial measures; for example, annual sales revenue, sales growth, annual profit and profit growth. Rather than seek empirical data, some studies ask managers for their perceptions; for example, their perception of sales growth compared to competitors.

Business Strategy: The main way the organization chooses to compete; for example, via cost leadership, differentiation, niche and so forth.

IT Alignment: The "fit" between the business and its IT; particularly, the fit between business strategy and IT strategy.

IT Implementation Success: Often, when a new system has been introduced, there is either a formal or informal evaluation of whether the system has benefited the organization. This evaluation could include the degree to which a system has achieved its expectations or goals.

IT Strategy: This refers to applications, technology and management; in particular, the IT applications an organization chooses to run, the IT technology it chooses to operate, and how the organization plans to manage the applications and the technology.

Small Firm: There is no universal definition. Most definitions are based on the number of employees, but some definitions include sales revenue. For example, 20 employees is the official definition in New Zealand, while in North America, a firm with up to 500 employees is defined as a small firm. Another important aspect of any definition of "small firm" is the firm's independence; that is, a small firm is typically considered to be independent, or not a subsidiary of another firm.

Information Technology and Virtual Communities

Chelley Vician
Michigan Technological University, USA

Mari W. Buche
Michigan Technological University, USA

INTRODUCTION AND BACKGROUND

Information technologies have made virtual communities possible. A community is a gathering of individuals who share something—be it knowledge, shared interests, a common purpose, or similar geographic surroundings. Traditionally, most communities are bound by time and space such that interaction and communication takes place in a same-time, same-place setting (Johansen, Sibbet, Benson, Martin, Mittman, & Saffo, 1991; Moffitt, 1999). The ready availability, high performance, and rapid diffusion of information technologies that enable communication across time, geography, and formal organizations now permits the development of communities that exist solely in the interaction activities made possible by IT (Igbaria, 1999). In essence, the community exists "virtually" through communication over the Internet (e.g., in cyberspace, as per Lee, Vogel, & Limayem, 2003) rather than taking on physical form at a specific time and in a specific geographic location.

There is little consensus among scholars and practitioners on a single definition of a virtual community (Lee et al., 2003), and several different terms are often used to label aspects of this phenomenon: online communities, communities of practice, virtual teams, e-learning, asynchronous learning networks, virtual classrooms, virtual learning, video-based information networks, discussion groups, and online forums. Table 1 provides representative sources for many of these alternative categorizations. However, shared characteristics of virtual communities are the following.

- Communication and interaction are primary activities of the community.
- Community interaction occurs through computer-mediated or computer-based communication.
- The content and process of the interaction is controlled by the community members.
- The community space is not geography or time bound, but is located in cyberspace through the networks and computers of individuals and the Internet.

Table 1. Virtual community examples

Type of Virtual Community	Source
community health and education	Gurstein (2000) Kodama (2001)
e-learning, asynchronous learning networks, virtual classrooms, virtual learning, learning community, online learning environment	DeSanctis, Fayard, Roach, & Jiang (2003) DeSanctis, Wright, & Jiang (2001) Hardaker & Smith (2002) Haynes & Holmevik (2001) Hiltz (1994) Hiltz & Wellman (1997) Holmevik & Haynes (2000) Piccoli, Ahmad, & Ives (2001)
online communities, communities of practice, virtual community	Blanchard & Markus (2004) Gurstein (2000) Rheingold (2000) Werry & Mowbray (2001) Williams & Cothrel (2000)
virtual teams	Lipnack & Stamps (2000) Powell, Piccoli, & Ives (2004) Townsend, DeMarie, & Hendrickson (1998)

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

This article will provide an overview of both the information technologies commonly used to sustain the virtual communities and representative examples of several kinds of virtual communities. Critical issues regarding the virtual-community phenomenon will also be presented.

INFORMATION TECHNOLOGIES

There is no single information technology but rather a convergence of several information technologies, the expansion of technology capacities, and human ingenuity in applying the burgeoning technological capabilities toward organizational and interpersonal uses that has precipitated the popularity of virtual communities. A virtual community exists because of the Internet and networks that enable the transmission and receipt of messages among people using computers for communication purposes. The most important technological components of a virtual community are (a) the Internet and the World Wide Web (WWW); (b) telecommunications and network hardware, software, and services; and (c) personal-computing hardware and software.

The Internet and the World Wide Web have evolved from specialized applications for scientists and researchers to a global information infrastructure easily accessed by the general public (Leiner et al., 2002). The Internet as we know it today owes its origins to ARPANET (wide-area network developed for the U.S. Defence Advanced Research Project Agency) and 1960s network researchers who were intent on proving the viability of connecting computers together to enable social interaction and communication (Leiner et al.). Today's Internet is a foundational "network of networks" that easily connects people worldwide with computing and communications technologies. The World Wide Web, in contrast, is a global hypertext system that uses the Internet as a means of providing information. Tim Berners-Lee (1998), inventor of the WWW concept and the first browser client and server in 1990, explains the difference between the Internet and the Web as follows:

The Web exists because of programs which communicate between computers on the Net. The Web could not be without the Net. The Web made the Net useful because people are really interested in information (not to mention knowledge and wisdom!) and don't really want to have to know about computers and cables.

Telecommunications and network hardware, software, and services provide the link between individual computers and the larger Internet capabilities[1]. Telecommunications and network hardware include routers and gateways that connect different networks and permit interoperability between different computers (Rowe, 2001). Advances in hardware capacities and capabilities (e.g., ready availability of broadband connections to the Internet) have facilitated the rapid diffusion of software applications that permit the sharing of data, voice, images, and video across the Internet. U.S. research shows that the number of broadband subscribers continues to increase over time (Webre, 2004), with asynchronous digital-subscriber-line (ADSL) connections growing at a rate comparable to cable modem connections (Federal Communications Commission, 2003). From a telecommunications services perspective, the growth of Internet service providers for the individual consumer and the expansion of networking groups within IT departments in organizations speak to the continuing importance of the network connection to the Internet.

Together with the network connections, ownership of personal-computing technologies continues to grow at a positive rate (Shiffler, 2004), which influences one's ability to join virtual communities. Current personal computers (PCs) now come equipped with more main memory, disk space, bus capacity, and processor speed than the largest mainframe computers used by network researchers in the 1960s (Laudon & Laudon, 2004). The PC's graphical user interface contributes to an individual's navigation of the operating system and software applications (Laudon & Laudon). Both the rate of PC ownership and the higher performance capacities of the PC units add to the increasing interest in virtual communities (Igbaria, 1999; Lee et al., 2003; Werry & Mowbray, 2001). PC software such as electronic mail and instant messaging also permit greater communication between individuals in cyberspace.

Virtual communities rely upon the reliable availability of the Internet, networking components, and personal-computing technologies to provide the space for individuals to congregate for a specific purpose.

The integration of these information technologies provides the means of connecting individuals without regard to geographic location and time of day. Clearly, the needs of human beings to meet people, join groups, and maintain associations with individuals having common interests now have ample technological means to sustain such social relationships in cyberspace.

EXAMPLES OF VIRTUAL COMMUNITIES

Prior to the advent of the World Wide Web, virtual communities were largely text-based adventures known as multiuser domains (MUDs) and multiuser-domain object oriented (MOO) (Holmevik & Haynes, 2000; Rheingold, 2002). MOOs continue to be used for educational purposes (Haynes & Holmevik, 2001), and the World Wide Web has made it easier for individuals to form virtual communities with interactive Web sites. Lee et al. (2003) report that out of a sample of 200 Web sites with some form of virtual community, 43% were relationship oriented, 38% were interest oriented, 12% were fantasy oriented, and 7% were transaction-based virtual communities.

More recently, businesses and organizations have begun experimenting with how to harness the capabilities of virtual communities to enhance operations and services (Williams & Cothrel, 2000). Health care or medically focused virtual communities continue to be a popular and successful experiment (Gurstein, 2000). Kaiser Permanente, a not-for-profit health maintenance organization (HMO), operates a successful virtual community focused on improving member services and promoting preventive health care. One of Kaiser's key success factors was the creation of an integrated, online environment such that members were empowered to make their own health-care decisions; a pilot study indicated improved customer satisfaction with the HMO (Williams & Cothrel). Governments have also begun experimenting with virtual communities as a means of providing health care and/or medical information to rural or outlying communities, especially as the multimedia technologies have improved their capabilities (Kodama, 2001).

Virtual communities are also well suited to bringing together individuals who might normally not be able to belong to a group due to diverse backgrounds, geographical distance, or time barriers. For example, independent contractors and consultants often work alone and in narrow specialties. The communication venues available in most virtual communities provide an independent consultant with the social networking that is vital to enhancing his or her own capabilities and services. About.com is a primary example of how a virtual community can be utilized to support such a distributed workforce (Williams & Cothrel, 2000). About.com permits each independent contractor (guide) to manage a Web site under the About.com umbrella on a particular topic (e.g., knitting, structured query language - SQL). About.com provides discussion forums, online training, and a resource area known as the "lounge" as support mechanisms for the independent contractors. Although the guides are not employees of About.com, they are managed as part of its larger workforce that provides information and entertainment services to the public. Two-way, computer-mediated communication has been central to the success of both the virtual community and About.com's management of freelance talent (Williams & Cothrel).

Virtual communities tend to be created for long-term objectives, while virtual teams often have a shorter life span by design (Lipnack & Stamps, 2000). Virtual teams have been used in both organizational and educational settings as a way of linking "groups of geographically, organizationally, and/or time-dispersed workers" (Powell et al., 2004, p. 7). Where a virtual community might have thousands of members, a virtual team often has less than 20 members. Variable makeup, dependence on computer-mediated communication, and capability to span both organizational boundaries and time restrictions are distinguishing characteristics of these teams (Powell et. al). Virtual teams, especially those formed for ad hoc or short-term reasons, provide an organization with high adaptability and flexibility in a competitive global marketplace (Maznevski & Chudoba, 2001) as specialized needs are recognized and acted upon. For organizations that wish to pilot test the features of virtual communities for operational reasons, the use of virtual teams can be an easy way to experiment with this new form of organizing and communicating.

A final example of successful virtual communities is the proliferation of learning environments. Again,

due to the capabilities of IT, learners can be connected within cyberspace when they cannot assemble in a single location that permits face-to-face interaction. Technology features support the learning objectives and provide for ample interaction among the participants. Many universities have begun utilizing such venues as part of their graduate education programs as computer-mediated communication enables the participation of global learners (DeSanctis et al., 2001; Hilsop, 1999; Hiltz & Wellman, 1997).

Current IT makes it simple to provide the space and place for the formation of virtual communities, though the actual sense of community is much harder to develop and sustain (Blanchard & Markus, 2004). This human element to the virtual community makes it possible for individuals to overcome time and location barriers such that lasting relationships can be formed (Walther, 1996). However, as with any human endeavor, not all attempts at virtual communities are successful. Unsuccessful or problematic virtual communities can and do occur: (a) Flaming and flame wars (e.g., generally negative and inflammatory electronic communication that would not normally be said if interacting in a face-to-face situation) can result due to the reduced social context in electronic communication (Alonzo & Aiken, 2004; Sproull & Kiesler, 1986); and (b) deviant, destructive behaviors in virtually constructed worlds may be seen more frequently than in reality (Powers, 2003; Suler & Phillips, 1998). Successful virtual communities have been able to leverage the features of IT to encourage positive human behaviors, while problematic virtual communities have struggled with managing the full range of human behaviors. Thus, there are ample opportunities for future research into the critical issues of human behavior in virtual communities.

CRITICAL ISSUES

Critical issues for virtual communities are centered around three main areas: (a) individual issues, (b) managerial issues, and (c) technological issues. Virtual community members must be able to communicate via computer-based communication tools and must have a comfort level with both the technologies and the communication activities (Lipnack & Stamps, 2000). Researchers are actively investigating the importance of trust (Jarvenpaa, Knoll, & Leidner, 1999; Siau & Shen, 2003; Suchan & Hayzak, 2001) and other individual-level factors such as computer-mediated communication anxiety (Brown, Fuller, & Vician, 2004) to understand their roles in e-based interaction inherent in virtual communities.

From a managerial perspective, there are many human-resource issues when organizations use virtual communities. Employee training, appraisal, and conflict management are but a few areas of concern (Williams & Cothrel, 2000). Additionally, there is the issue of how to make the virtual community experience one in which the employees will want to participate. A virtual community must have participant interaction, and if employees will not participate, the virtual community will have a difficult time getting started and maintaining itself (Blanchard & Markus, 2004; Williams & Cothrel).

The major technological issues have to do with the continuing improvements in capacity, features, and kinds of information technologies by manufacturers. As IT continues to evolve, individuals and organizations will need to stay abreast of the technological developments and determine the best ways to leverage the new capabilities in the virtual communities of the future. Today's version of cyberspace that requires text-based input (either from keyboard or keypad) may soon be eclipsed by voice input integrated with images and wearable computing devices (Jennings, 2003). Virtual communities that can take advantage of future technological developments will continue to thrive.

CONCLUSION

Virtual communities have evolved rapidly based on human needs and the opportunities created by integrated networks. As innovations continue to be developed, the number of virtual communities and teams will likely increase. In addition, the personal experiences of the participants and their sense of presence within virtual teams will improve. The most important benefit of these technologies is the ability of individuals to communicate, collaborate, and cooperate without regard to separation due to time and space. The Internet and the World Wide Web have managed to make the planet a much smaller place for networked individuals.

REFERENCES

Alonzo, M., & Aiken, M. (2004). Flaming in electronic communication. *Decision Support Systems, 36*(3), 205-213.

Berners-Lee, T. (1998). *Frequently asked questions by the press—Tim BL. General questions 1998: What is the difference between the Net and the Web?* Retrieved June 10, 2004, from *http://www.w3.org/People/Berners-Lee/FAQ.html #InternetWeb*

Blanchard, A. L., & Markus, L. M. (2004). The experienced "sense" of a virtual community: Characteristics and processes. *Database, 35*(1), 65-79.

Brown, S. A., Fuller, R. M., & Vician, C. (2004). Who's afraid of the virtual world? Anxiety and computer-mediated communication. *Journal of the Association for Information Systems, 5*(2), Article 3. Retrieved March 31, 2004, from *http://jais.isworld.org/articles/default.asp?vol=5&art=3*

DeSanctis, G., Fayard, A., Roach, M., & Jiang, L. (2003). Learning in online forums. *European Management Journal, 21*(5), 565-577.

DeSanctis, G., Wright, M., & Jiang, L. (2001). Building a global learning community. *Communications of the ACM, 44*(12), 80-82.

Federal Communications Commission. (2003). *High-speed services for Internet access: Status as of June 30, 2003.* Retrieved June 1, 2004, from *http://www.fcc.gov/wcb/stats*

Free on-line dictionary of computing. (n.d.). Retrieved from *http://wombat.doc.ic.ac.uk/foldoc/index.html*

Gurstein, M. (2000). *Community informatics: Enabling communities with information and communications technologies*. Hershey, PA: Idea Group Publishing.

Hardaker, G., & Smith, D. (2002). E-learning communities, virtual markets, and knowledge creation. *European Business Review, 14*(5), 342-350.

Haynes, C., & Holmevik, J. R. (Eds.). (2001). *High wired: On the design, use, and theory of educational MOOs* (2nd ed.). Ann Arbor, MI: University of Michigan Press.

Hilsop, G. W. (1999). Anytime, anyplace learning in an online graduate professional degree program. *Group Decision and Negotiation, 8*(5), 385-390.

Hiltz, S. R. (1994). *The virtual classroom: Learning without limits via computer networks*. Norwood, NJ: Ablex.

Hiltz, S. R., & Wellman, B. (1997). Asynchronous learning networks as a virtual classroom. *Communications of the ACM, 40*(9), 44-49.

Holmevik, J. R., & Haynes, C. (2000). *MOOversity: A student's guide to online learning environments*. New York: Pearson Education Longman.

Igbaria, M. (1999). The driving forces in the virtual society. *Communications of the ACM, 42*(12), 64-70.

Jarvenpaa, S., Knoll, K., & Leidner, D. (1999). Communication and trust in global virtual teams. *Organization Science, 10*(6), 791-815.

Jennings, L. (2003). From virtual communities to smart mobs. *The Futurist, 37*(3), 6-8.

Johansen, R., Sibbet, D., Benson, S., Martin, A., Mittman, R., & Saffo, P. (1991). *Leading business teams*. New York: Addison-Wesley.

Kodama, M. (2001). New regional community creation, medical and educational applications through video-based information networks. *Systems Research and Behavioral Science, 18*, 225-240.

Laudon, K. C., & Laudon, J. P. (2004). *Management information systems* (8th ed.). Upper Saddle River, NJ: Prentice Hall.

Lee, F. S. L., Vogel, D., & Limayem, M. (2003). Virtual community informatics: A review and research agenda. *The Journal of Information Technology Theory and Application (JITTA), 5*(1), 47-61.

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., et al. (2002). *All about the Internet: A brief history of the Internet.* Internet Society (ISOC). Retrieved May 31, 2004, from *http://www.isoc.org/internet/history/brief.shtml*

Lipnack, J., & Stamps, J. (2000). *Virtual teams: People working across boundaries with technology*. New York: John Wiley & Sons.

Maznevski, M., & Chudoba, K. (2001). Bridging space over time: Global virtual team dynamics and effectiveness. *Organization Science, 11*(5), 473-492.

Moffitt, L. C. (1999). A complex system named community. *Journal of the Community Development Society, 30*(2), 232-242.

Piccoli, G., Ahmad, R., & Ives, B. (2001). Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic IT skills training. *MIS Quarterly, 25*(4), 401-426.

Powell, A., Piccoli, G., & Ives, B. (2004). Virtual teams: A review of current literature and directions for future research. *Database, 35*(1), 6-36.

Powers, T. M. (2003). Real wrongs in virtual communities. *Ethics and Information Technology, 5*, 191-198.

Rheingold, H. (2000). *The virtual community: Homesteading on the electronic frontier.* Boston, MA: The MIT Press. Available online at *http://www.rheingold.com/vc/books/*

Rowe, S. H., II. (2001). *Telecommunications for managers* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Shiffler, G. (2004). *Forecast: PCs, worldwide and United States, March 2004 update* (Executive summary). Stamford, CT: Gartner Group.

Siau, K., & Shen, Z. (2003). Building customer trust in mobile commerce. *Communications of the ACM, 46*(4), 91-94.

Sproull, L. S., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science, 32*(11), 1492-1513.

Suchan, J., & Hayzak, G. (2001). The communication characteristics of virtual teams: A case study. *IEEE Transactions on Professional Communication, 44*(3), 174-186.

Suler, J. R., & Phillips, W. (1998). The bad boys of cyberspace: Deviant behavior in multimedia chat communities. *Cyberpsychology and Behavior, 1*, 275-294.

Townsend, A., DeMarie, S., & Hendrickson, A. (1998). Virtual teams: Technology and the workplace of the future. *Academy of Management Executive, 12*(3), 17-29.

Walther, J.B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Researc, 23*(1), 3-43.

Webre, P. (2004). *Is the United States falling behind in adopting broadband? Congressional Budget Office economic and budget issue brief.* Retrieved June 10, 2004, from *http://www.cbo.gov/briefs.cfm*

Werry, C., & Mowbray, M. (Eds.). (2001). *Online communities*. Upper Saddle River, NJ: Prentice Hall.

Williams, R. L., & Cothrel, J. (2000, Summer). Four smart ways to run online communities. *Sloan Management Review, 41*(4), 81-91.

KEY TERMS

Bandwidth: The difference between the highest and lowest frequencies of a transmission channel (the width of its allocated band of frequencies).

Baud: The unit in which the information-carrying capacity or signaling rate of a communication channel is measured. One baud is one symbol (state transition or level transition) per second.

Broadband: A class of communication channels capable of supporting a wide range of frequencies, typically from audio up to video frequencies. A broadband channel can carry multiple signals by dividing the total capacity into multiple, independent bandwidth channels, where each channel operates only on a specific range of frequencies. The term has come to be used for any kind of Internet connection with a download speed of more than 56K baud.

Browser: A software program running on a client computer that allows a person to read hypertext. The browser permits viewing the contents of pages and navigating from one page to another. Netscape Navigator, Microsoft Internet Explorer, and Lynx are common browser examples.

Digital Subscriber Line (DSL): A family of digital telecommunications protocols designed to allow high-speed data communication over the existing copper telephone lines between end users and telephone companies.

Hypertext: A collection of documents containing cross-references that, with the aid of a browser program, allow the reader to move easily from one document to another.

Integrated Services Digital Network (ISDN): A set of communications standards allowing a single wire or optical fibre to carry voice, digital network services, and video that may replace the plain, old telephone system.

Internet: The Internet is the largest network in the world. It is a three-level hierarchy composed of backbone networks, midlevel networks, and stub networks. These include commercial (.com or .co), university (.ac or .edu), other research networks (.org, .net), and military (.mil) networks, and they span many different physical networks around the world with various protocols, chiefly the Internet protocol.

Internet Service Provider (ISP): A company that provides other companies or individuals with access to, or presence on, the Internet.

Modem (Modulator/Demodulator): An electronic device for converting between serial data from a computer and an audio signal suitable for transmission over a telephone line or cable TV wiring connected to another modem.

MUD Object Oriented (MOO): One of the many MUD spin-offs created to diversify the realm of interactive, text-based gaming. A MOO is similar to a MUSH in that the users themselves can create objects, rooms, and code to add to the environment.

Multiuser Dimension/Multiuser Domain (MUD): Originally known as multiuser dungeons, MUDs are a class of multiplayer, interactive games that are text-based in nature and accessible via the Internet or a modem. A MUD is like a real-time chat forum with structure; it has multiple "locations" like an adventure game and may include combat, traps, puzzles, magic, or a simple economic system. A MUD where characters can build more structure onto the database that represents the existing world is sometimes known as a MUSH (multiuser shared hallucination). MUDs originated in Europe and spread rapidly around the world.

Network: Hardware and software data-communication systems that permit communication among computers and the sharing of peripheral devices (e.g., printers).

Protocol: A set of formal rules describing how to transmit data, especially across a network.

Router: A device that forwards packets (messages) between networks.

World Wide Web (WWW): An Internet client-server hypertext distributed information-retrieval system that originated from the CERN High-Energy Physics laboratories in Switzerland.

Integrated Platform for Networked and User-Oriented Virtual Clothing

Pascal Volino
University of Geneva, Switzerland

Thomas Di Giacomo
University of Geneva, Switzerland

Fabien Dellas
University of Geneva, Switzerland

Nadia Magnenat-Thalmann
University of Geneva, Switzerland

INTRODUCTION

Fashionizer is an integrated framework that fits the needs of the garment industry of virtual garment design and prototyping, concentrating on simulation and visualization features.

Virtual Try On has been developed in close relationship to be compliant with Fashionizer's clothes and to allow trying them virtually on a body's *avatar* in real time on the Web; in a few words, it is a virtual clothing boutique.

The framework integrates innovative tools aimed for efficiency and quality in the process of garment design and prototyping, taking advantage of state-of-the-art algorithms from the field of mechanical simulation, computer animation, and rendering that are directly provided by the research team of MIRALab.

APPROACH AND RESEARCH

To take a 2-D (two-dimensional) pattern as a base is the simplest way to obtain a precise, exact, and measurable description of a 2-D surface, which is the representative of the virtual fabric. In the traditional clothing industry, one garment is composed of several 2-D surfaces (pattern pieces) that need to be seamed together in a particular way to describe the complete garment. Fashionizer enables clothes designers to create 3-D (three-dimensional) clothes based on pat-

Figure 1. An example of 2-D patterns applied on a body with Fashionizer

Figure 2. Different points of view for viewing the worn garment

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

terns. Users will be able to alter the patterns in the 2-D view and visualize automatically the simulated garment in the 3-D view.

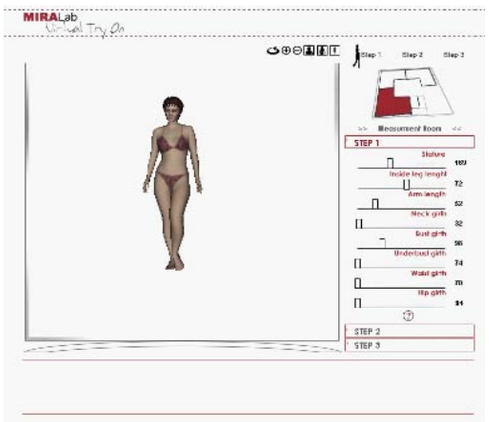
It also allows the user to dress virtual humans with realistic simulated clothes, based on the designed patterns, and therefore to simulate and display the final aspect of the garment, in dynamic situations as well, before manufacturing it. Through built-in plug-ins, patterns can be imported from traditional CAD systems, or can be created manually. Furthermore, 3-D generic models of bodies, female or male, are manipulated and crafted based on anthropomorphic measurements.

Fashionizer provides functionality from the most recent research, namely, physical and realistic simulation of fabrics; that is, each kind of woven fabric can be simulated with respect to its texture, thinness, and properties of textile. The simulation of clothes is based on the *finite elements method* that provides the most accurate and precise results (Volino & Magnenat-Thalmann, 2001). Fashionizer also provides less accurate methods based on *mass-spring systems* from research done for more interactive simulations (Volino & Magnenat-Thalmann, 1997). Moreover, Fashionizer can animate a whole sequence of simulated clothes, which involves a robust simulation of clothes and efficient collision detections between clothes and the underlying body (Volino & Magnenat-Thalmann, 2000a, 2000b). This accuracy provides an estimation of pressure and stretching areas on the body that is wearing the simulated cloth in order to measure and visualize the comfort and fitting of a garment on a specific body.

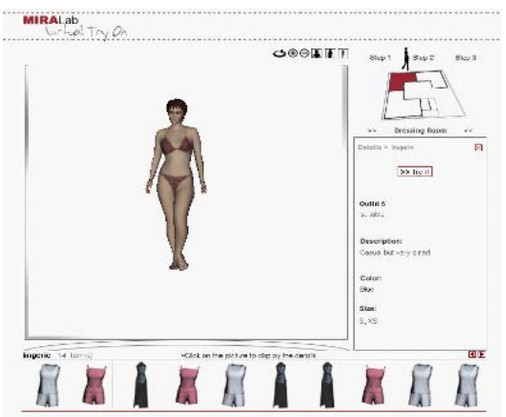
The Real Time Virtual Try On is an altogether new approach to online visualization and immersion that lets any standard Web browser display interactive 3-D dressed virtual bodies. Our approach provides a minimal response time to the user since a major part of the content to be manipulated is generated on the client side rather than on the server. The MIRALab Virtual Try On client application is not only involved in the visualization of garments, but also used for the calculation of the cloth and body deformation. The question is "What is needed for virtually trying on clothes in real time?" First, a virtual copy of the user's body measurements and a database of virtual clothes to be tried are required, and finally, a real-time display of the whole is mandatory to illustrate how the cloth fits and reacts in real time.

Figure 3. From top right to bottom left are the three steps of our Virtual Try On

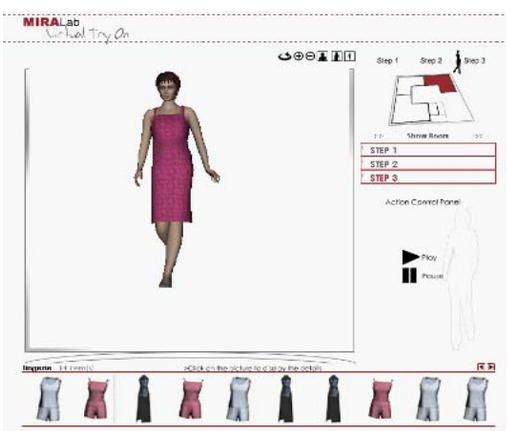
- First the user loads the avatar according to his/her body measurements



- Then the user selects a desired cloth



- Finally the user can have a look at the moving cloth on his/her avatar



First the user loads the avatar according to her or his body measurements. Then the user selects a desired cloth. Finally the user can have a look of the moving cloth on her or his avatar.

The generation of an avatar based on a user's personal measurements is another challenging issue for computer-graphics research: In fact, it is inconceivable to store each different body in a database because the amount of data is huge and would not fit for Web applications. The answer is provided by parameterized shape modifications (Magnenat-Thalmann, Seo, & Cordier, 2003; Seo, Cordier, & Magnenat-Thalmann, 2003) and implemented in the Virtual Try On. By taking a set of extreme types of body, appropriate and evolved *interpolations* between them generate a body specifically to a user's measurements. The main characteristics of each initial body are extracted by a *principal-component analysis*, helpful for the generation of new individualized bodies.

Visualizing and simulating efficiently the cloth worn on the user's moving avatar is another important issue. Precomputed sequences of walking animations can be stored, but cloth movements are too complicated and dynamic to be precomputed off line. Actually, the simulation does not need to be physically accurate since what is interesting for the potential consumer is to have a true aspect of the cloth on his or her body before buying it. Thus, the simulation should only be plausible visually while physical accuracy is an optional bonus. Following this assumption, the cloth model is simplified, in terms of polygons, to simulate garments in real time. The method is based on a statistical learning of cloth's movement behaviour and on a segmentation of the cloth in three layers: loose, tight, and middle parts. For further details see Cordier and Magnenat-Thalmann (2002) and Cordier, Seo, and Magnenat-Thalmann (2003).

ACKNOWLEDGEMENTS

The authors would like to thank Christiane Luible, Hyewon Seo, and Frederic Cordier for their development, consulting, and help.

REFERENCES

Cordier, F., & Magnenat-Thalmann, N. (2002). Real-time animation of dressed virtual humans. *Eurographics Conference Proceedings,* July (pp. 327-336).

Cordier, F., Seo, H., & Magnenat-Thalmann, N. (2003, January/February). Made-to-measure technologies for online clothing store. *IEEE Computer Graphics and Applications*, *23*(1), 38-48.

Magnenat-Thalmann, N., Seo, H., & Cordier, F. (2003). Automatic modeling of virtual humans and body clothing. *Proceedings of 3-D Digital Imaging and Modeling*, October (pp. 2-10).

Seo, H., Cordier, F., & Magnenat-Thalmann, N. (2003, July). Synthesizing animatable body models with parameterized shape modifications. *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 120-125.

Volino, P., & Magnenat-Thalmann, N. (1997). Developing simulation techniques for an interactive system. In *Proceedings of the 1997 International Conference on Virtual Systems and MultiMedia,* Washington, D.C. (pp. 1-9). IEEE Computer Society.

Volino, P., & Magnenat-Thalmann, N. (2000a, May). *Accurate collision response on polygonal meshes.* Computer Animation Conference, Philadelphia, PA.

Volino, P., & Magnenat-Thalmann, N. (2000b, June). Implementing fast cloth simulation with collision response. In *Proceedings of the Internatioinal Conference on Computer Graphics,* Washington, D.C. (p. 257). IEEE Computer Society.

Volino, P., & Magnenat-Thalmann, N. (2001). Comparing efficiency of integration methods for cloth animation. *Proceedings of Computer Graphics International (CGI)*, July (pp. 265-274).

KEY TERMS

Avatar: A virtual representation generated by computers. It can be, for example, a copy of a user's body to try on virtual clothes.

Finite Elements Method: A second approach to simulate soft bodies and deformations. It is also used to model fabrics by considering its surface as a continuum and not a fixed set of points. This method is more accurate but slower to compute.

Interpolation: A family of mathematical functions to compute unknown states between two known states. For instance, it is possible to interpolate between two 3-D models of a body to obtain an intermediate one.

Mass-Spring System: A set of particles linked by springs. Each particle is characterized by a 3-D position and a mass, and is linked to its neighbours by springs (with their own physical properties). This method can simulate the different existing mechanical interactions of a deformable object.

Principal-Component Analysis: A mathematical method based on statistics to extract the main "behaviours" of a set of data.

Interactive Digital Television

Margherita Pagani
Bocconi University, Italy

BACKGROUND

Interactive television (iTV) can be defined as the result of the process of convergence between television and the new interactive digital technologies (Pagani, 2000, 2003).

Interactive television is basically domestic television boosted by interactive functions that are usually supplied through a back channel. The distinctive feature of interactive television is the possibility that the new digital technologies can give the user the ability to interact with the content that is on offer (Flew, 2002; Owe, 1999; Pagani, 2000, 2003).

The evolution toward interactive television has not just an exclusively technological, but also a profound impact on the whole economic system of digital broadcaster—from offer types to consumption modes, and from technological and productive structures to business models.

This article attempts to analyze how the addition of interactivity to television brings fundamental changes to the broadcasting industry.

This article first defines interactive transmission systems and classifies the different services offered according to the level of interactivity determined by two fundamental factors such as response time and return channel band.

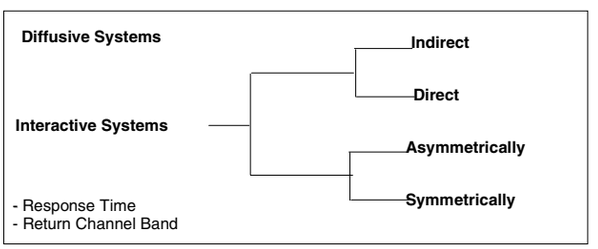
After defining the conceptual framework and the technological dimension of the phenomenon, the article analyzes the new types of interactive services offered.

The Interactive Digital Television (iDTV) value chain will be discussed to give an understanding of the different business elements involved.

A DEFINITION OF INTERACTIVITY

The term *interactivity* is usually taken to mean the chance for interactive communication among subjects (Pagani, 2003). Technically, interactivity implies the presence of a return channel in the communication system, going from the user to the source of information. The channel is a vehicle for the data bytes that represent the choices or reactions of the user (input).

Table 1. The classification of communication systems



This definition classifies systems according to whether they are diffusive or interactive (Table 1).

- Diffusive systems are those that only have one channel that runs from the information source to the user (this is known as *downstream*);
- Interactive systems have a return channel from the user to the information source (this is known as *upstream*).

There are two fundamental factors determining performance in terms of system interactivity: response time and return channel band.

The more rapidly a system's response time to the user's actions, the greater is the system's interactivity. Systems thus can be classified into:

- *Indirect interactive systems* when the response time generates an appreciable lag from the user's viewpoint;
- *Direct interactive systems* when the response time is either very short (a matter of a few seconds) or is imperceptible (real-time).

The nature of the interaction is determined by the bit-rate that is available in the return channel. This can

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

allow for the transfer of simple impulses (yes—no logic), or it can be the vehicle for complex multimedia information (i.e., in the case of videoconferencing). From this point of view, systems can be defined as asymmetrically interactive when the flow of information is predominantly downstream. They also can be defined as symmetrical when the flow of information is equally distributed in the two directions (Huffman, 2002).

Based on the classification of transmission systems above previously, multimedia services can be classified into diffusive (analog or digital) and interactive (Table 2).

Digital television can provide diffusive numerical services and asymmetrical interactive video services. Services such as videoconferencing, telework, and telemedicine, which are within the symmetrical interactive video based upon the above classification, are not part of the digital television offers.

Local Interactivity

An interactive application that is based on local interactivity is commonly indicated as «enhanced TV» application. It does not require a return-path back to the service provider.

An example is the broadcaster transmitting a football match using a «multi-camera angle» feature, transmitting the video signals from six match cameras simultaneously in adjacent channels. This allows the viewer to watch the match from a succession of different vantage points, personalizing the experience. One or more of the channels can be broadcast within a time delay for instant replays.

This application involves no signal being sent back to the broadcaster to obtain the extra data. The viewer is simply dipping in and out of that datastream to pick up supplemental information as required.

One-Way Interactivity

One-way interactivity refers to all interactive applications in which the viewer did send back a signal to the service provider via a return path, but there is no ongoing, continuous, two-way, real-time dialogue, and the user doesn't receive a personalized response.

The most obvious application is direct response advertising. The viewer clicks on an icon during a TV commercial (if interested in the product), which sends a capsule of information containing the viewer's details to the advertiser, allowing a brochure or sample to be delivered to the viewer's home.

Two-Way Interactivity

Two-way interactivity is what the technological purist defines as «true» interactivity. The user sends data to a service provider or other user, which travels along a return path, and the service provider or user sends data back, either via the return path itself or «over the air». Two-way interactivity presupposes

Table 2. Classes of service (classes not directly relevant to interactive multimedia services are in grey)

Class of services	Services (examples)
1. DIFFUSIVE SERVICES	
Analogue transmission	Free channels, Pay TV
Numerical diffusion	Digital channels Pay Per View (PPV) Near Video On Demand (NVOD)
2. INTERACTIVE SERVICES	
Asymmetric interactive video	Video On Demand (VOD), Music On Demand, TV Shopping, Interactive advertising Interactive games, TV banking
Low speed data	Telephony (POTS), data at 14,4; 28,8; 64; 128 Kbit/s
Symmetric interactive video	Co-operative work, Tele-work, Tele-medicine, Videoconference, Multi-videoconference
High speed data	Virtual reality, distribution of real time applications

«addressability»—the senders and receivers must be able to address a specific dataset to another sender or receiver.

What might be termed «low level» two-way interactivity is demonstrated by a TV pay-per-view service. Using the remote control, the viewer calls up through an on-screen menu a specific movie or event scheduled for a given time and «orders» it. The service provider than ensures, by sending back a message to the viewer's set top box, that the specific channel carrying the movie at the time specified is unscrambled by that particular box, and that that particular viewer is billed for it.

Low-level two-way interactivity is characterized by the fact that the use of the return path back to the service provider is peripheral to the main event.

«High level» two-way interactivity, on the other hand, is characterized by a continuing two-way exchange of data between the user and the service provider (i.e. video-conferencing, Web surfing, multiplayer gaming, and communications-based applications such as chat and SMS messaging).

INTERACTIVE TELEVISION

Interactive television can be defined as domestic television boosted by interactive functions, made possible by the significant effects of digital technology on television transmission systems (ETSI, 2000; Flew, 2002; Nielsen, 1997; Owen 1999). It supports subscriber-initiated choices or actions that are related to one or more video programming streams (FCC, 2001; Pagani, 2003).

A first level of analysis shows that interactive television is a system through which the viewer can ask something to the program provider. In this way, the viewer can transmit his or her own requests through the two-way information flow, made possible by the digitalization of the television signal.

The viewer's reception of the digital signal is made possible through a digital adapter (set top box or decoder), which is connected to the normal television set or integrated with the digital television in the latest versions. The set top box decodes the digital signals in order to make them readable by the conventional analogue television set (Figure 1). The set top box has a memory and decoding capacity that allows it to handle and visualize information. Thus, the viewer can accede to a simple form of interactivity by connecting the device to the domestic telephone line. In addition, other installation and infrastructure arrangements are required, depending on the particular technology. In particular, a return channel must be activated. This can imply a second dedicated telephone line for return path via

Table 3. Interactive television services

Category	Interactive application
Enhanced Tv	Personalized weather information Personalized EPG (Electronic Program Guide) Menu à la carte Different viewing angles Parental Control Enhanced TV Multi language choice
Games	Single player games Multiplayer games Voting and Betting
Communication	Instant messages E-mail
Finance	Financial information Tv Banking
E-commerce	Pay Per View TV Shopping
Advertising	Interactive Advertising
Internet	Web access

modem. The end user can interact with his or her TV set through a special remote control or, in some cases, even with a wireless keyboard.

TYPES OF INTERACTIVE TV SERVICES

The British broadcasting regulator Independent Television Commission (ITC) differentiates between two essentially different types of interactive TV services: dedicated and program-related.

- *Program-related services* refer to interactive TV services that are directly related to one or more video programming streams. These services allow users to obtain additional data related to the content (either programming or advertising), to select options from a menu, to play or bet along with a show or sports event, or to interact with other viewers of the same program.
- *Dedicated services* are stand-alone services not related to any specific programming stream. They follow a model closer to the Web, even if there are differences in hyperlinks, media usage, and, subsequently, mode of persuasion. This type of interactive service includes entertainment, information, and transaction services.

Interactive TV services can be classified further into some main categories (Table 3).

PROGRAM-RELATED SERVICES

Electronic Program Guide (EPG)

EPG is a navigational device allowing the viewer to search for a particular program by theme or other category and order it to be displayed on demand.

EPG helps people grasp a planning concept, understand complex programs, absorb large amount of information quickly, and navigate in the TV environment.

Typical features are:

- **Flip:** Displaying the current channel, the name of the program, and its start and end time.
- **Video Browser:** Allows viewers to see program listings for other channels.
- **Multi-Language Choice**.
- **VCR Programming**.

More advanced features under development concern:

- **Customization:** Displaying features like favorites or reminders, which can be set for any future program.
- **Ranking Systems:** Seen as preference systems, where viewers can order channels, from the most watched to the least watched.
- **Noise Filters:** Seen as systems in which viewers block information (i.e., removing channels that they never watch). One related issue is parental control (filter), where objectionable programming can be restricted by setting locks on channels, movies, or specific programs.

Pay Per View

Pay per view services provide an alternative to the broadcast environment; through broadband connections, they offer viewers on-demand access to a variety of server-based content on non-linear basis. Viewers pay for specific programs.

DEDICATED SERVICES

Interactive Games

Interactive game shows take place in relation to game shows, to allow viewers to participate in the game. Network games allow users to compare scores and correspond by a form of electronic mail, or to compete against other players.

There are different revenue models related to the offer of games: subscription fee, pay-per play or pay per day, advertising, sponsorship, banner.

Interactive Advertising

Interactive advertising is synchronized with a TV ad. An interactive overlay or icon is generated on the screen, leading to the interactive component. When the specific pages are accessed, viewers can learn

more about products, but generally, other forms of interactions also are proposed. Viewers can order catalogues; benefit from a product test; and participate in competition, draw, or play games.

The interactive ad should be short in order not to interfere with the program that viewers wish to watch. The message must be simple and quick. This strategy is based on provoking an impulsive response (look at the interactive ad) resulting in the required action (ordering the catalogue). A natural extension of this concept is to enable consumers to order directly.

TV Shopping

TV shopping is common both on regular channels and on specialized channels. Some channels are specialized in teleshopping (i.e., QVC and Home Shopping Europe). Other channels develop interactive teleshopping programs (i.e., TF1 via TPS in France). Consumers can order products currently shown in the teleshopping program and pay by inserting their credit card in the set-top box card reader. During the program, an icon appears, signaling viewers that they can now buy the item.

The chosen product is then automatically displayed in the shopping basket. Viewers enter the quantity and the credit card number. The objectives of such programs are to give viewers the feeling of trying products. The products' merits are demonstrated in every dimension allowed by the medium. In some ways, we can consider teleshopping as the multimedia counterpart missing from Web shops.

Mixing elements of teleshopping and e-commerce might constitute a useful example of integration of TV and interactivity, resulting in a new form of interactive shopping. Consumers can be enticed by attractive features and seductive plots.

There is a difference between interactive advertising and interactive shopping. Initially, interactive advertising is triggered from an ad and concerns a specific product. Shops, on the other hand, are accessed directly from the TV shopping section and concern a range of products.

TV shopping presents a business model close to PPV and has a huge potential.

TV Banking

TV banking enables consumers to consult their bank statements and carry out their day-to-day banking operations (financial operations, personalized investment advice, or consult the Stock Exchange online).

Interactive TV gives financial service companies a new scope for marketing; it permits them to display their products in full-length programs rather than commercials lasting a few seconds and to deliver financial advice in interactive formats, even in real time. Such companies particularly value the ability to hot-link traditional TV commercials to sites where viewers can buy products online. In addition, service providers on interactive TV can tailor their offers precisely by collecting detailed data about the way customers use the medium.

Designing online services for TV requires video and content development skills that few banks have in-house, requiring them, in all likelihood, to join forces with television and media specialists.

Table 4. The iDTV value chain: players and added value

Player	Added Value
Content provider	Produce content edit/format content for different iDTV platform
Application Developer	Research and develop interactive applications
Content aggregator	Acquire content rights, reformat, package and rebrand content
Network operator	Maintain and operate network, provide adequate bandwidth
iDTV Platform operator	- Acquire aggregated content and integrate into iDTV service applications - Host content/outsource hosting - Negotiate commerce deals - Bundle content/service into customer packages - Track customer usage and personalize offering
Customer equipment	- Research and development equipment - Manufacture equipment - Negotiate deals and partnerships

INTERACTIVE DIGITAL TELEVISION (iTV) VALUE CHAIN

The interactive digital television marketplace is complex, with competing platforms and technologies providing different capabilities and opportunities.

The multi-channel revolution, coupled with the developments of interactive technology, is truly going to have a profound effect on the supply chain of the TV industry.

The competitive development generated by interactivity creates new business areas, requiring new positioning along the value chain for existing operators.

Several types of companies are involved in the iDTV business: content provider, application developer, broadcasters, network operator, iDTV platform operator, hardware and software developer, Internet developers also interested in developing for television, consultants, research companies, advertising agencies, etc. (Table 4).

A central role is played by broadcasters whose goal is to acquire contents from content providers (banks, holders of movie rights, retailers), store them (storage), and define a broadcast planning system (planning). They directly control users' access as well as the quality of the service and its future development (Figure 1).

Conditional access is an encryption/decryption management method (security system) through which the broadcaster controls the subscriber's access to digital and iTV services, such that only those authorized can receive the transmission. Conditional access services currently offered include, other than encryption/decryption of the channel, also security in purchase and other transactions, smart card enabling, and issuing and customer management services (billing and telephone servicing). The subscriber most often uses smart cards and a private PIN number to access the iTV services. Not all services are purchased necessarily from the conditional access operator.

Service providers, such as data managers, provide technologies that allow the broadcaster to deliver personalized, targeted content. They use Subscriber Management System (SMS) to organize and operate the company business. The SMS contains all customer-relevant information and is responsible for keeping track of placed orders, credit limits, invoicing and payments, as well as the generation of reports and statistics.

Satellite platforms, cable networks, and telecommunications operators mainly focus on the distribution of the TV signal, gradually tending to integrate upstream in order to have a direct control over the production of interactive services.

Figure 1. The iDTV value chain: Head-end phases

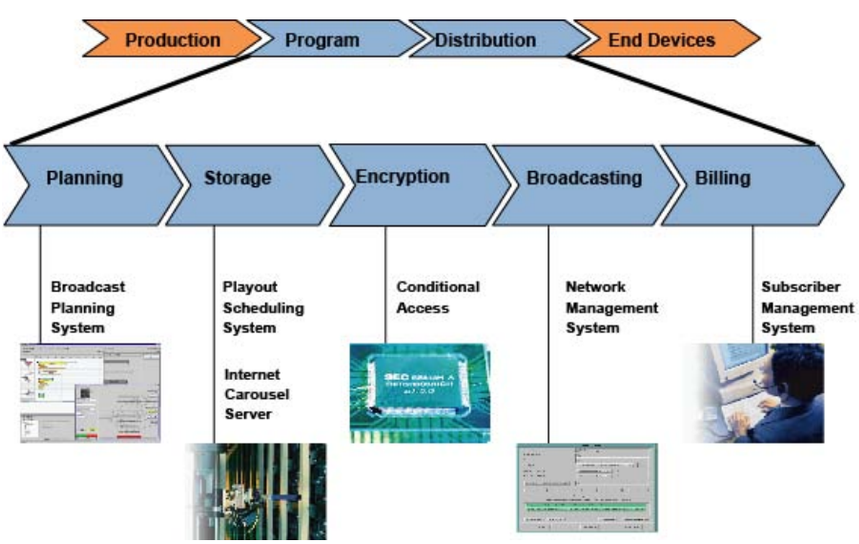
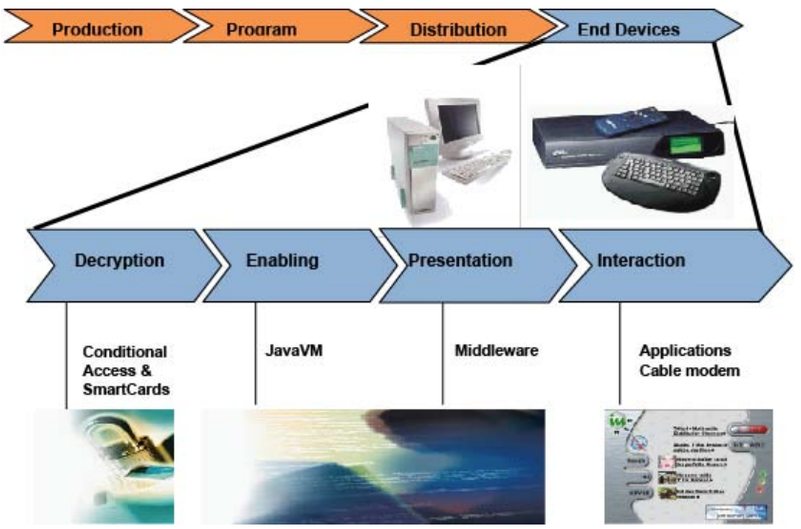


Figure 2. The iDTV value chain: End device



The vast end device segment (Figure 2) includes two subsegments regarding the hardware and the software embedded in it.

The hardware manufacturers (e.g., Sony, Philips, Nokia, etc.) design, produce, and assembly the set-top boxes (STB).

The software subsegment includes:

1. Operating systems developers (i.e., Java Virtual Machine by SunMicrosystem, Windows CE by Microsoft, and Linus) provide many services, such as resource allocation, scheduling, input/output control, and data management. Although operating systems are predominantly software, partial or complete hardware implementations may be made in the form of firmware.
2. Middleware providers and developers provide programming that serves to glue together or mediate between two separate and usually already-existing programs. Middleware in iTV is also referred to the Application Programming Interface (API); it functions as a transition/conversion layer of network architecture that ensures compatibility between the basal infrastructure (the operating system) and diverse upper-level applications. There are four competing technologies: Canal+ Media Highway (running on Java OS); Liberate Technologies (Java); Microsoft TV (Windows CE); and OpenTV (Spyglass). These are all proprietary solutions acting as technological barriers trying to lock-in the customers. This situation creates vertical market where there is no interoperability, and only programs and applications written specifically for a system can run on it.
3. User-level applications provider includes interactive gaming, interactive (or electronic) programming guides, Internet tools (e-mail, surfing, chat, instant messaging), t-commerce, video-on-demand (VOD) and personal video recording (PVR).

CONCLUSION

Interactive TV services are providing welcome opportunities for brand marketers who are keen to pursue closer relationships with a more targeted audience, with the promise of a new direct sales channel complete with transactional functionality. For broadcasters, garnering marketer support and partners can be a crucial means of reducing costs, providing added bite to marketing digital TV to consumers, while establishing new sources of revenue (based on carriage fees from advertisers,

revenue shares for transactions coordinated via the digital TV platform, and payment for leads generation and data accrued through direct marketing).

From a strategic point of view, the main concern for broadcasters and advertisers will be how to incorporate the potential for interactivity, maximizing revenue opportunities and avoiding the pitfalls that a brand new medium will afford. It is impossible to offer solutions, merely educated guesses for how interactive TV will develop.

Success will depend upon people's interests in differentiated interactive services.

1. First, the development of a clear consumer proposition is crucial in a potentially confusing and crowded marketplace.
2. Second, the provision of engaging, or even unique, content will continue to be of prime importance.
3. Third, the ability to strike the right kind of alliances is a necessity in a climate that is spawning mergers and partnerships. Those who have developed a coherent strategy for partnering with key companies that can give them distribution and content naturally will be better placed.
4. Finally, marketing the service and making it attractive to the consumer will require considerable attention, not to mention investment.

In summary, the development of the market generated by technological innovations forces the individual television firm to know increasingly its positioning and the state of the dynamic competition.

REFERENCES

Bowler, J. (2000). DTV content exploitation. What does it entail and where do I start? *New TV Strategies, 2*(7), 7.

Datamonitor. (2001). *Is the channel dead? The impact of interactivity on the TV industry.* Datamonitor Report.

ETSI. (2000). Digital video broadcasting (DVB); Interaction channel for satellite distribution systems. ETSI EN 301 790 V1.2.2 (2000-12).

Flynn, B. (2000). *Digital TV, Internet & mobile convergence—Developments and projections for Europe*. Digiscope Report. London: Phillips Global Media.

FCC. (2001). *In the matter of non-discrimination in the distribution of interactive television service over cable.* CS Docket, No. 01-7, 2.

Flew, T. (2002). *New media: An introduction*. Melbourne: Oxford University Press.

Grebb, M. (2002). The power of cable and telecommunications. *Multichannel News, 9*, 14.

Huffman, F. (2002). Content distribution and delivery. *Proceedings of the 56th Annual NAB Broadcast Eng. Conference,* Las Vegas, Nevada.

Nielsen, J. (1997). TV meets the Web. Retrieved on January 3, 2004 from *http://www.useit.com/alertbox/9701.html*

Owen, B. (1999). *The Internet challenge to television*. Cambridge, MA: Harvard University Press.

Pagani, M. (2000). Interactive television: A model of analysis of business economic dynamics. *Journal of Media Management JMM, 2*(1), 25-37.

Pagani, M. (2000). *Interactive television: The managerial implications* [working paper]. Milan, Italy: I-LAB Research Center On Digital Economy – Bocconi University.

Pagani, M. (2001). Le implicazioni manageriali delle nuove tecnologie digitali interattive sul broadcaster televisivo. *Proceedings of the Conference SISEI, EGEA*, Milan, Italy.

Pagani, M. (2003). *Multimedia and interactive digital TV: Managing the opportunities created by digital convergence*. Hershey, PA: Idea Publishing Group.

Rawolle, J., & Hess, T. (2000). New digital media and devices: An analysis for the media industry. *Journal of Media Management JMM, 2*(II), 89-98.

KEY TERMS

Broadband: A network capable of delivering high bandwidth. Broadband networks are used by Internet and cable television providers. For cable, they range from 550 MHz to 1GHz. A single TV regular broadcast channel requires 6MHz, for example. In the Internet domain, bandwidth is measured in bits-per-second (BPS).

Decoder: See Set-Top Box.

Interactive Television: Can be defined as domestic television boosted by interactive functions, made possible by the significant effects of digital technology on television transmission systems. It supports subscriber-initiated choices or actions that are related to one or more video programming streams.

Interactivity: Usually taken to mean the chance for interactive communication among subjects. Technically, interactivity implies the presence of a return channel in the communication system, going from the user to the source of information. The channel is a vehicle for the data bytes that represent the choices or reactions of the user (input).

Multimedia Service: Refers to a type of service, which includes more than one type of information (text, audio, pictures, and video) transmitted through the same mechanism and allowing the user to interact or modify the information provided.

Set-Top Box: The physical box that is connected to the TV set and the modem/cable return path. It decodes the incoming digital signal, verifies access rights and security levels, displays cinema-quality pictures on the TV set, outputs digital surround sound, and processes and renders the interactive TV services.

Value Chain: As made explicit by Porter in 1980, a value chain can be defined as a firm's co-coordinated set of activities to satisfy customer needs, starting with relationship with suppliers and procurement, going through production, selling and marketing, and delivering to the customer. Each stage of the value chain is linked with the next stage and looks forward to the customer's needs and backwards from the customer, too. Each link of the value chain must seek competitive advantage; it must be either a lower cost than the corresponding link in competing firms, or it must add more value by superior quality or differentiated features (Koch, 2000).

Interactive Memex

Sheng-Uei Guan
National University of Singapore, Singapore

INTRODUCTION

With the development of the Internet, a great deal of information is on-line. Popular search sites could be visited million times daily and the sites related to your interest will often be visited by you. Although bookmarks can be used to record frequented Web sites, browsers discard most history and trail information. The explosion of information needs a more effective mechanism. Memex has been considered in this domain. Assisted by Memex, a Web surfer can retrieve the URL trails that a user visited several months ago. In this paper, we propose a mechanism Self-modifiable Color Petri Net - SCPN to simulate the Memex functions in a Web browser. In this mechanism, an SCPN instance is used to record a trail of a topic, a place in an SCPN instance represents a Web site.

RELATED WORK

Petri Net

Petri Net is a graphical notation for the formal description of systems whose dynamics are characterized by concurrency, synchronization, mutual exclusion, and other conflict, which are typical features of distributed environment. A formal definition of Petri Nets is a four-tuple (P, T, I, O) (Peterson, 1981) where P is a set of places that are the state variables of a system; T is a set of transitions, which are state changing operators. I and O are the pre- and post-conditions of a transition. The dynamic performance of a Petri Net is controlled by the firing rule.

Several extended Petri Net models have been proposed to extend its application domains. Examples of which are Object Composition Petri Net (OCPN) in (Little, 1990) and Enhanced Prioritized Petri Net (EP-net) in (Guan, 1999) and (Guan, 2002) which is an enhanced version of P-net in (Guan, 1998). The general concepts of Petri Net are described in the next section. Self-modifiable Color Petri Net (SCPN) is also introduced in the next section.

Memex

As early as 1945, Vannevar Bush proposed a desktop personal information machine called the Memex (memory extender) (Bush, 1945). Memex focused on the problems of "locating relevant information in the published records and recording how that information is intellectually connected". An important feature of Memex is the function of associative indexing that presents the feature of hyperlinks. In addition to these links, Bush also wanted Memex to support the building of trails through the material in the form of a set of links that would combine information of relevance for a specific topic.

Some Powerful Bookmarks, Bookmark Organizers, and Other Works

There are quite a number of powerful bookmarks and organizers developed like the Personal Web Map (PWM) (Yamada, 1999), Bookmark Organizer (Maarek, 1996), PowerBookmarks (Li, 1999), and CZWeb (Fisher, 1997). All of these provide organization and management of bookmarks but not Memex functions, that is, they do not provide surfing history and trails.

A related work which uses trails is Memoir (Derource, 2001). Trails are used to open hypermedia link services and a set of software agents to assist users in accessing and navigating vast amounts of information in Intranet environments. The trails in Memoir are mainly used to record actions on documents that users have visited. In our Memex application, trails are mainly used to record and retrieve surfing history information.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

PETRI NET AND SELF-MODIFIABLE COLOR PETRI NET (SCPN)

A Petri Net structure, P, is a four-tuple.

$P = (P, T, I, O)$

i. $P = \{p_1, p_2, \ldots p_x\}$, where $x \geq 0$, is a finite set of *Places*.
ii. $T = \{t_1, t_2, \ldots t_y\}$, where $y \geq 0$, is a finite set of *Transitions*.

where $P \cap T = \varnothing$, i.e., the set of the places and transitions are disjoint.

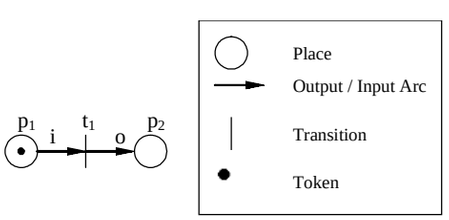
iii. $I: T \rightarrow P^{\infty}$ is the *Input Arc*, a mapping from places to bags of transitions.
iv. $O: T \rightarrow P^{\infty}$ is the *Output Arc*, a mapping from transitions to bags of places.

$Token = \{token_1, token_2, \ldots token_x\}$, $x \geq 0$, $x \in \Im$, is a finite set of dynamic markings on places.

The Petri Net model consists of places, transitions, arcs, and tokens.

i. A *place*, denoted by a circle, represents the state of the system. p_1 and p_2 in Figure 1 are places.
ii. A *transition*, denoted by a vertical line, represents the action of the system and is led by an output arc and trailed by an input arc. t_1 in Figure 1, led by *o* and trailed by *i*, is a transition.
iii. An *arc* represents the flow relation between transitions and places.
iv. An *input arc*, denoted by an arc terminated by an arrowhead leading from a place to a transition, maps a place to a transition. *i* in Figure 1 is an input arc.

Figure 1. Petri Net segment



v. An *output arc*, denoted by an arc terminated by an arrowhead leading from a transition to a place, maps a transition to a place. *o* in Figure 1 is an output arc.
vi. A *token* is a marking that denote the current state of the system. A firing of a transition removes a token from its input place and places a token in its output place. In Figure 1, a token is marked in place, p_1.
vii. The *input place of a transition* is the place that is connected to the transition via an input arc.
viii. The *output place of a transition* is the place that is connected to the transition via an output arc.

The Petri Net is governed by a set of Firing Rules that allows movement from one state to another.

i. A transition is *enabled* when all input places that are connected to it via an input arc have at least one token.
ii. A firing of a transition removes a token from its input place and places a token in its output place.

Introducing some novel mechanisms to Petri Net gives birth to SCPN which can handle user interaction flexibly. Unlike in Petri net, SCPN has two types of tokens: color tokens and resource tokens. Resource token are divided into two sub-types: a forward token that moves in the same direction with arcs and a reverse token that moves in the opposite direction with arcs. In SCPN, certain commands for each mechanism are also introduced.

For the new mechanisms to work, some new rules are defined to assist SCPN to complete its functions:

- A color token will be injected into each place that contains resource token(s) when a user interaction occurs.
- When a color token is injected, the execution of the model will be interrupted.
- When all the commands associated with a color token have been executed, this color token will be deleted. Then the playback of resource tokens will be resumed.

The commands associated with each color token can be designed according to the corresponding user interaction. In the following, we use some solid examples to demonstrate how color tokens are used to realize Memex functions in Web surfing.

DESIGNING MEMEX FUNCTIONS USING SCPN

Simulating Memex Trail Recording in Web Browsing

To simulate Memex in Web browsing, we assume that a place in SCPN represents a Web site. Each time a Web site is opened, a color token including the following basic commands will be injected into the place p_{start} that includes a resource token as shown in Figure 2: lock the resource token in p_{start}, create a new place p_1 (this place will represent the newly opened Web site), create a new transition t_1, create an arc from current place p_{start} to the new transition t_1, create an arc from the new transition t_1 to the new place p_1, unlock the resource token in $p_{start.}$ Finally, the color token self-deletes, transition t_1 fires, the resource token moves to p_1 indicating that the Web site represented by this place is active now. While SCPN is recording the surfing trail, the corresponding Web site address will be recorded along with each place.

Main Trail and Side Trails

Almost all Web sites contain some related hyperlinks. A trail can bifurcate: when a hyperlink of one Web site is visited, a side-trail will be created to record it. As shown in Figure 3, the main trail that represents the main surfing history is composed by places with *m* as the first subscript, the side-trail that represents the hyperlink of a Web site is composed by places with names having *s* as the first subscript.

If the hyperlink is opened in a new window or the user wants to record the hyperlink of a Web site as a new trail, a new starting place will be created as the first place in a new trail as shown in Figure 4. The arcs linking from p_{m1} to $p_{m'1}$ are represented by dot lines meaning that these arcs do not allow a reverse token moving along them.

Backward and Forward Operations

Using SCPN to record a browser trail, it can simulate the *backward* and *forward* operations of Web browsing. A resource token in a place indicates that the Web site corresponding to this place is active, the arcs indicate the sequence of Web sites being visited. When a user issues a *backward* command, a color token corresponding to this command will be injected into the place p_{m3} that includes a resource token as shown in Figure 5a. Then the commands associated with this color token executes, the resource token in p_{m4} is locked and changed to a reverse one as shown in Figure 5b. In Figure 5c, the reverse token is unlocked and the color token self-deletes. Finally, transition t_{m3} fires, the reverse token moves from p_{m3} to p_{m2} and changes back to forward resource token as shown in Figure 5d, the information of the Web site related to place p_{m2} will be retrieved. At the same time, p_{m3} is recorded as an exit point so that a future forward move will allow p_{m3} to be revisited.

Figure 2. An Example: Using SCPN to record the surfing trail

a. Create a new place and transition

b. Transition t_1 fires

Figure 3. Trail recording using SCPN

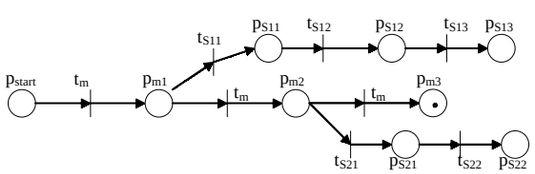


Figure 4. A new trail is created

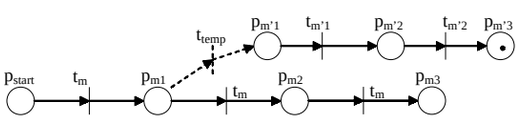
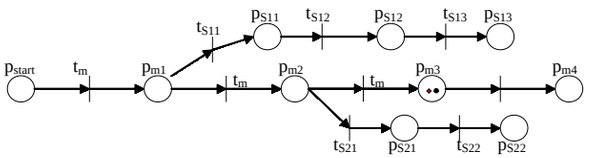
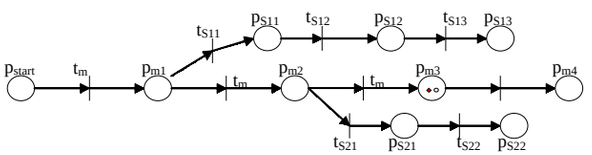


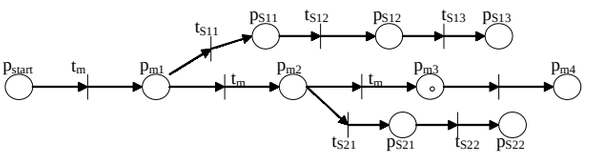
Figure 5. Implementation of the backward operation



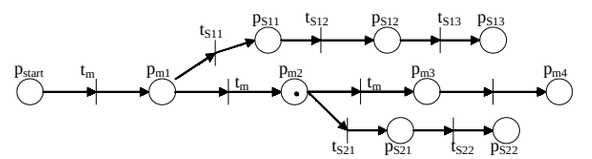
a. A color token corresponding to the backward operation is injected into p_{m3}.



b. The resource token is locked and changed to a reverse one.



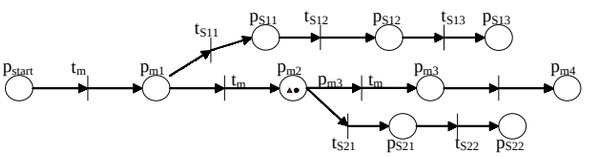
c. Reverse token is unlocked and color token self-deletes.



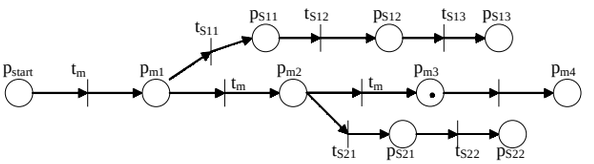
d. Execution of the backward operation is completed.

After checking the content of this Web site, if the user decides to go back to the previous Web site again, a *forward* command can be issued. A color token associated with the *forward* command will be injected into place p_{m2} that contains the resource token as shown in Figure 6a. Then the command executes to direct the resource token to fire. At this moment, we can see that one of the two transitions t_{m3} and t_{s21} can fire. In modeling Memex functions, SCPN is used to

Figure 6. Implementation of the forward operation



a. A color token associated with the forward operation is injected into p_{m2}.



b. The forward operation executes.

record the surfing history, the resource token is used to indicate the active Web site. There can be only one place that can contain the resource token at a time. In such a *forward* operation, because the exit point of a previous *backward* operation has been recorded, t_{m3} will fire and the resource token will move to p_{m3} as shown in Figure 6b, at the same time, the record of the previous exit point will be replaced by p_{m2} for future use.

SIMULATOR

Using Visual C++, a simulator has been built. This simulator can model Memex functions such as trail recording and retrieval. To make the simulation more realistic, we use the Microsoft Active X® controller in our program to display a Web site visited at the same time when the SCPN place corresponding to the Web site is created or a resource token is injected into the place. A user can click the buttons as shown in Figure 7 to simulate the corresponding function.

To make the simulator more powerful, a basic Petri Net design tool is provided. A Petri Net instance can be designed simply by clicking and dragging the icons from the toolbar to the white area. The Petri Net instance created can be saved as a

.mex file for future use. Also a *RUN* button as shown in the menu in Figure 7 is provided to execute a Petri Net instance. When the *RUN* button is clicked, the Petri Net instance will be executed. If the instance is active, the token will move according to the firing direction. The *Back, Forward* and *History* buttons are used to simulate Memex functions in a Web Browser. The *Save* button is used to save the trail. If some trails have been built, *Search* function can help a user to find an item of interest in these trails. We give an example to show how this simulator works.

As shown in Figure 8, when a Web site is opened (assume this is a new trail to be built), an event signal will be sent to the system indicating a new Web site is opened. With this event, a place will be created to record it. And a resource token will be created in the place at the same time to indicate that the Web site corresponding to this place is active. In order to let the user arrange trails according to his need, the simulator provides trail recording options. Each time when a Web site is opened, a dialog box will be popped up to ask the user if the Web site needs to be recorded as shown in Figure 9. If the user chooses not to archive this Web site, the place being created to record this Web site will be deleted after the Web site is closed. If the user puts down an existing trail name, the Web site will be added and recorded as the last place in this existing trail. If the user puts down a new trail name, a dialog box will be popped up to let the user choose how to record this Web site as shown in Figure 10. For example, if a hyperlink is followed after three Web sites have been visited, the user chooses to record it as a side trail by clicking the *Yes* button (Figure 10). This Web site will then be recorded as a side trail as shown in Figure 11.

As shown in Figure 12, there are five places in the SCPN instance shown. From this we know that five Web sites have been visited. The active Web site is http://www.google.com associated with the fifth place. SCPN can show how many Web sites have been visited and which one is active now, but no detailed information of these Web sites is shown on

Figure 7. The user interface of the Memex simulator

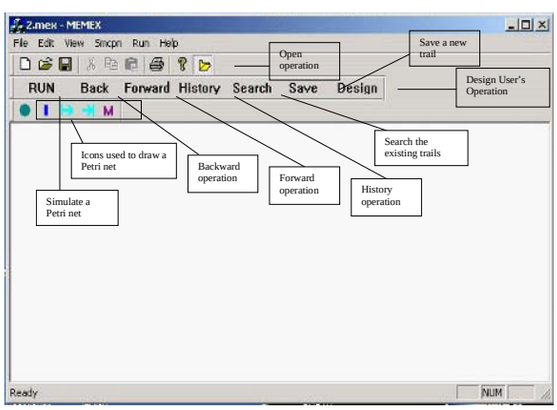


Figure 8. A place created to record the Web site being opened

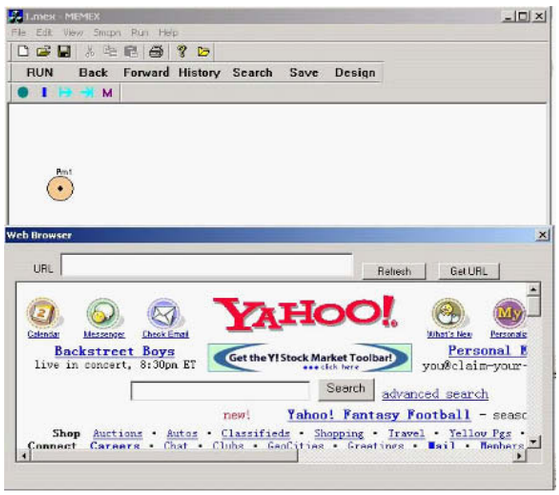


Figure 9. Archiving choice dialog box

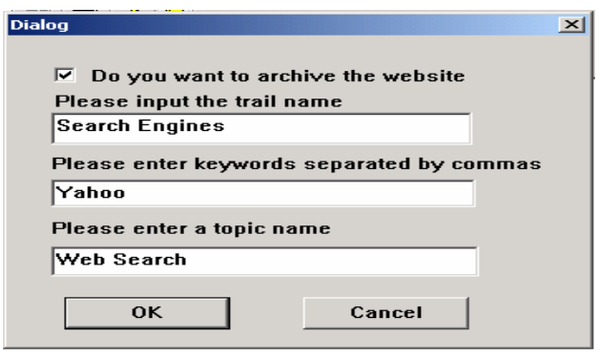


Figure 10. Trail creation choice dialog box

Trail Select Dialog
The input trail name does not exist, do you want to record it as a side trail (click 'Yes'), a new trail (click 'No') or continue to use the current trail (click 'Cancel')?
Yes
No
Cancel

the graph. If the user wants to see the details of the Web sites visited, the *History* button in the menu can accomplish this task.

Using SCPN to record trails, each place is associated with a Web site. It is easy to display history records. When the user issues a command '*History*', this can be done by clicking on the *History* button, a dialog box will be opened to show the detailed trail information as shown in Figure 12.

With the trail shown, we can select any item to revisit. For example, if we want to visit the IEEE Xplore Web site, just select it from the list and click the *ok* button. The corresponding Web site will be retrieved and the resource token will move to p_{m4} as shown in Figure 13.

In addition to trail recording and retrieval, the Memex simulator can also achieve the *backward* and *forward* operations similar to those functions in Web browsers. As shown in Figure 14, if the user wants to visit the previous Web site before the IEEE Xplore Web site, he only needs to click the *Back* button. The resource token will move to p_{m3} and at the same time the Web site associated with this place will be opened.

Following the above example, if the user wants to visit the next Web site again, he only needs to click the *Forward* icon, the resource token will move to p_{m4} and the corresponding Web site will be reopened at the same time.

Besides these Web-browser-like operations, the most important Memex function is that when some trails have been built, a user can search for it according to name/topic/keyword. As shown in Figure 15, when a user clicks the *Search* button, a dialog box will be popped up to show the existing trails. Then the user can select from these trails the one that he is interested in to retrieve or input it in the search Edit-box. For example, if the user wants to find some information

Figure 11. A Web site recorded as a side trail

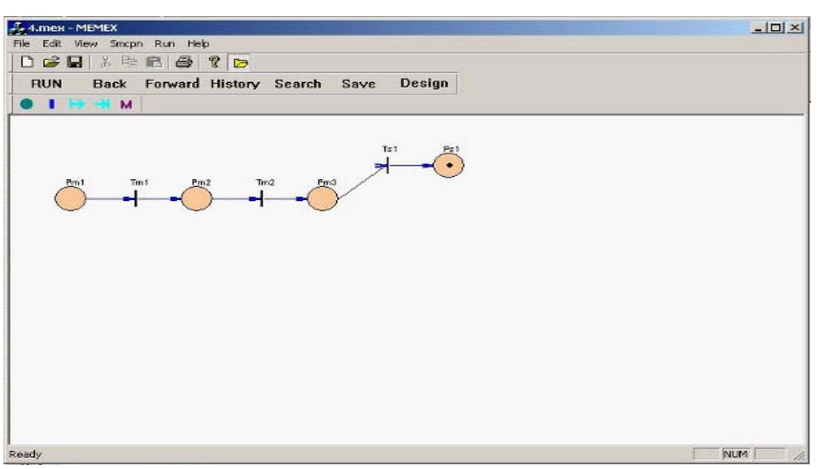


Figure 12. The history information displayed

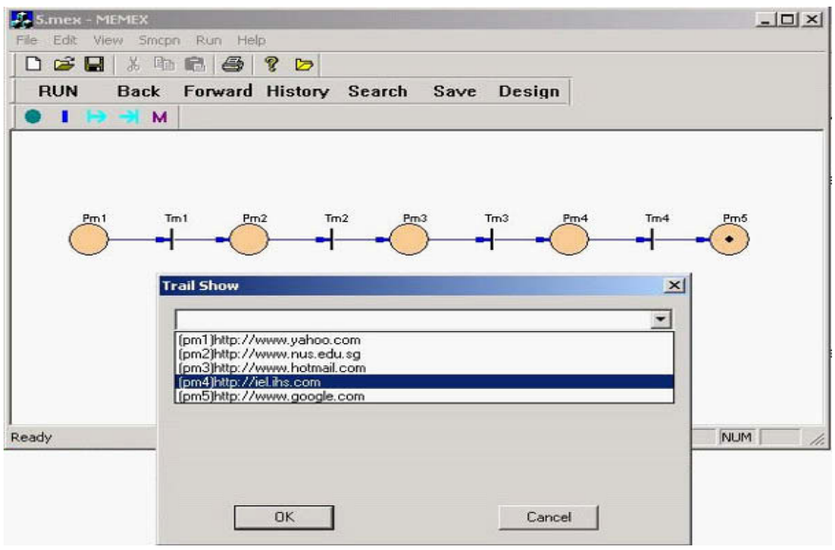


Figure 13. Web site represented by p_{m4} retrieved

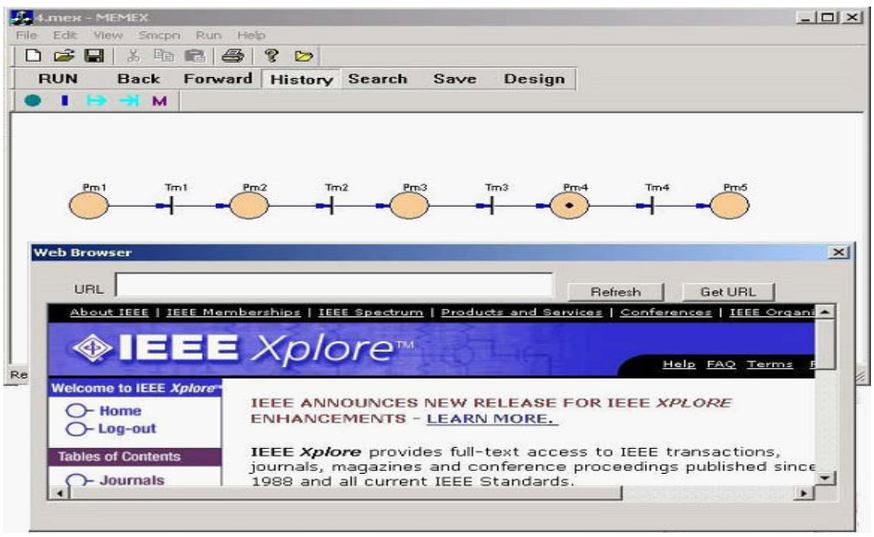

4.mex - MEMEX
File Edit View Smcpn Run Help
RUN Back Forward History Search Save Design
Pm1 Tm1 Pm2 Tm2 Pm3 Tm3 Pm4 Tm4 Pm5
Web Browser
URL
Refresh
Get URL
About IEEE | IEEE Memberships | IEEE Spectrum | Products and Services | Conferences
IEEE Xplore™
Help FAQ Terms
Welcome to IEEE Xplore
Home
Log-out
Tables of Contents
Journals
IEEE ANNOUNCES NEW RELEASE FOR IEEE XPLORE ENHANCEMENTS - LEARN MORE.
IEEE Xplore provides full-text access to IEEE transactions, journals, magazines and conference proceedings published since 1988 and all current IEEE Standards.

Figure 14. The backward operation executed

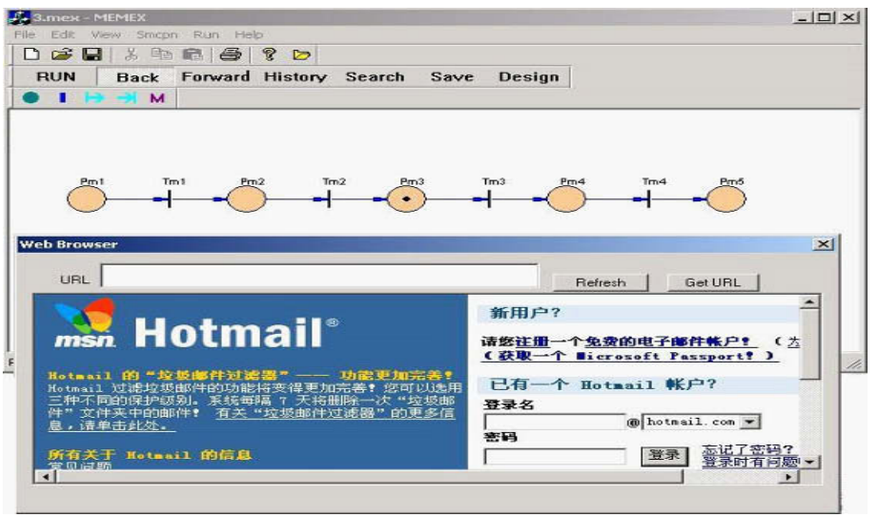
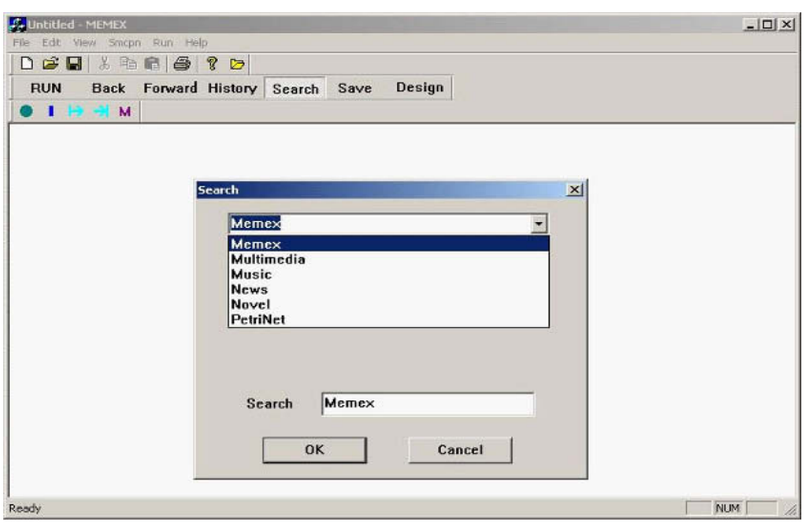


Figure 15. Search for existing trails by name

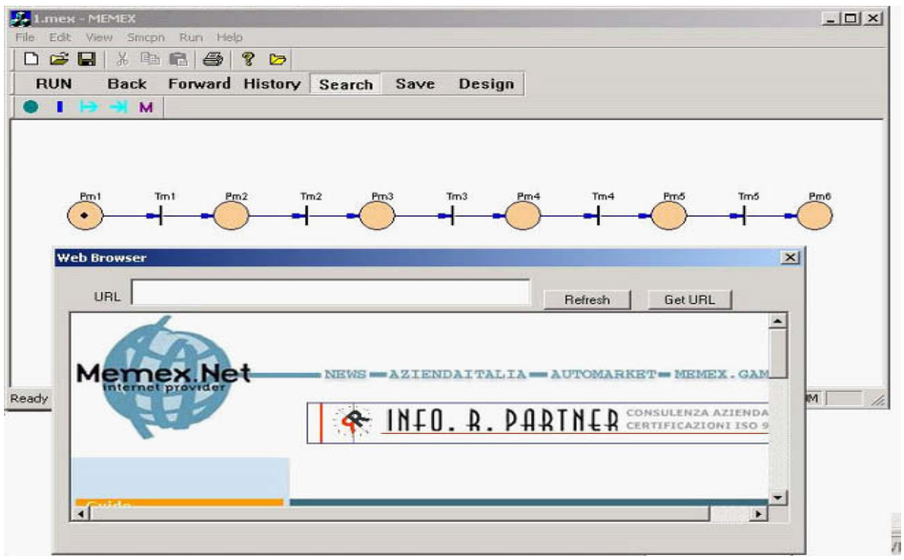


about Memex, he/she only needs to select the first item from the dialog box or input Memex into the search Edit-box and click the *ok* button. The Memex trail details will then be displayed in a dialog box. Then the user proceeds to choose a Web site he wants to visit from this trail. Assume that the first one is selected, the Web site will be opened and at the same time the trail represented by SCPN is displayed as shown in Figure 16. The resource token in p_{m1} indicates that the Web site associated with this place is active.

CONCLUSION

In this paper, we have given an introduction to a Self-modifiable Color Petri Net model - SCPN. With the powerful reconfiguration function offered from this

Figure 16. The first Web site of the Memex trail retrieved



model, Memex functions can be achieved in Web browsing. Our approach offers an underlying model with which a systematic approach to constructing Memex-like applications can be adopted. A simulator with user-friendly interface has been built to show how this can be achieved. This simulator can also be used as a Petri Net design tool to help users to design and implement their own Self-modifiable Color Petri Net instances.

REFERENCES

Al-Salqan, Y. & Chang, C. (1996). Temporal telations and synchronization agents. *IEEE Multimedia, 3*, 30-39.

Bulterman, D.C.A. (2002). SMIL 2.0.2. Examples and comparisons. *IEEE Multimedia, 9*(1), 74-84.

Bush, V. (1945). As we may think. *Atlantic Monthly, 176*, 101-108. Online at *http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm*

Chakrabarti, S., Srivastave, S., Subramanyam, M. & Tiwari, M. (2000). Using memex to archive and mine community Web browsing experience. *Computer Networks, 33*(1-6), 669-684. Online at *http://www9.org/w9cdrom/98/98.html*

Derource, D., Hall, W., Reich, S., Hill, G., Pikrakis, A. & Stairmand, M. (2001). Memoir: An open framework for enhanced navigation of distributed information. *Information Processing & Management, 37,* 53-74.

Fisher, B., Agelidis, G., Dill, J., Tan, P., Collaud, G., & Jones, C. (1997). CZWeb: Fish-eye views for visualiziang the world-wide web. *Proceedings of the Seventh International Conference on Human-Computer Interaction (HCI International '97)*, (pp. 719-722).

Guan, S. & Lim, S. (2002). Modeling multimedia with enhanced prioritized Petri Nets. *Computer Communications,* (8), 812-824.

Guan, S. & Lim, S. (1999). An enhanced prioritized Petri Net model for authoring interactive multimedia applications. *Proceedings the Second International Conference on Information, Communications & Signal Processing (ICICS'99)*, Singapore.

Guan, S., Yu, H., & Yang, J. (1998). A prioritized Petri Net model and its application in distributed

multimedia systems. *IEEE Transactions on Computers,* (4), 477-481.

Jensen, K. (1997). *Coloured Petri nets* (Vol. 1). Springer-Verlag.

Li, W.-S., Vu, Q., Agrawal, D., Hara, Y., & Takano, H. (1999). PowerBookmarks: A system for personalizable Web information organization, sharing and management. *Computer Networks, 31*, 1375-1389.

Little, T. & Ghafoor, A. (1990). Synchronization and storage models for multimedia objects. *IEEE Journal on Selected Area in Communication,* (3), 413-427.

Maarek, Y.S. & Ben Shaul, I.Z. (1996). Automatically organizing bookmarks per content. In *Proceedings Fifth International World Wide Web Conference*, Paris. Online at *http://www5conf.inria.fr/fich_html/papers/P37/Overview.html*

Peterson, J.L. (1981). *Petri net theory and the modeling of systems.* NJ: Prentice-Hall.

Yamada, S. & Nagino, N. (1999). Constructing a personal Web map with anytime-control of Web robots. *CoopIS'99 Proceedings, IFCIS International Conference on Cooperative Information Systems*, (pp. 140-147).

KEY TERMS

Distributed Environment: An environment in which different components and objects comprising an application can be located on different computers connected to a network.

History Retrieval: In Web browsing, it is the act of recalling Web sites that have been previously visited.

Modeling: The act of representing something (usually on a smaller scale).

Petri Nets: A directed, bipartite graph in which nodes are either "places" (represented by circles) or "transitions" (represented by rectangles), invented by Carl Adam Petri. A Petri Net is marked by placing "tokens" on places. When all the places with arcs to a transition (its input places) have a token, the transition "fires", removing a token from each input place and adding a token to each place pointed to by the transition (its output places). Petri Nets are used to model concurrent systems, particularly network protocols.

Synchronization: In multimedia, synchronization is the act of coordinating different media to occur or recur at the same time.

Tokens: An abstract concept passed between places to ensure synchronized access to a shared resource in a distributed environment.

Trail: In this work, it refers to a track of Web sites that have been visited.

User interaction: In multimedia, this is the act of users intervening or influencing in multimedia presentation.

Interactive Multimedia Technologies for Distance Education in Developing Countries

Hakikur Rahman
SDNP, Bangladesh

INTRODUCTION

With the extended application of information technologies (IT), the conventional education system has crossed physical boundaries to reach the un-reached through a virtual education system. In the distant mode of education, students get the opportunity for education through self-learning methods with the use of technology-mediated techniques. Accumulating a few other available technologies, efforts are being made to promote distance education in the remotest regions of developing countries through institutional collaborations and adaptive use of collaborative learning systems (Rahman, 2000a).

Distance education in a networked environment demands extensive use of computerized Local-Area and Wide-Area Networks (LAN/WAN), excessive use of bandwidth and expensive use of sophisticated networking equipment; in a sense this has become a hard-to-achieve target in developing countries. High initial investment cost always demarcates thorough usage of networked hierarchies where the basic backbone infrastructure of IT is in a rudimentary stage.

Developed countries are taking a leading role in spearheading distance education through flexible learning methods, and many renowned universities of the western world are offering highly specialized and demanding distance education courses by using their dedicated high-bandwidth computer networks. Many others have accepted a dual mode of education rather than sticking to the conventional education system. Research indicates that teaching and studying at a distance can be as effective as traditional instruction when the method and technologies used are appropriate to the instructional tasks with intensive learner-to-learner interactions and instructor-to-learner interactions. Radio, television and computer technologies, including the Internet and interactive multimedia methods, are major components of virtual learning methodologies.

The goals of distance education, as an alternative to traditional education, have been to offer accredited education programs, to eradicate illiteracy in developing countries, to provide capacity-development programs for better economic growth, and to offer curriculum enrichment in a non-formal educational arena. Distance education has experienced dramatic global growth since the early 1980s. It has evolved from early correspondence learning using primarily print-based materials into a global movement using various technologies.

BACKGROUND

Distance education has been defined as an educational process in which a significant proportion of the teaching is conducted by someone removed in space and/or time from the learner. Open learning, in turn, is an organized educational activity based on the use of teaching materials, in which constraints on study are minimized in terms either of access, or of time and place, pace, method of study or any combination of these (UNESCO, 2001).

There is no ideal model of distance education, but several are innovative for very different reasons. Philosophies of an approach to distance education differ (Thach & Murphy, 1994). With the advent of educational technology-based resources (CD-ROMs, the Internet, Web pages, etc.), flexible learning methodologies are getting popular to a large mass of the population who otherwise was missing the opportunity of accessing formal education (Kochmer, 1995). Murphy (1995) reported that to reframe the quality of teaching and learning at a distance, four types of interaction are necessary: learner-content, learner-teacher, learner-learner and learner-interface. Interaction also represents the connectivity the students feel with their professor, aides, facilitators and peers (Sherry, 1996). Responsibility for this

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

sort of interaction mainly depends upon the instructor (Barker & Baker, 1995).

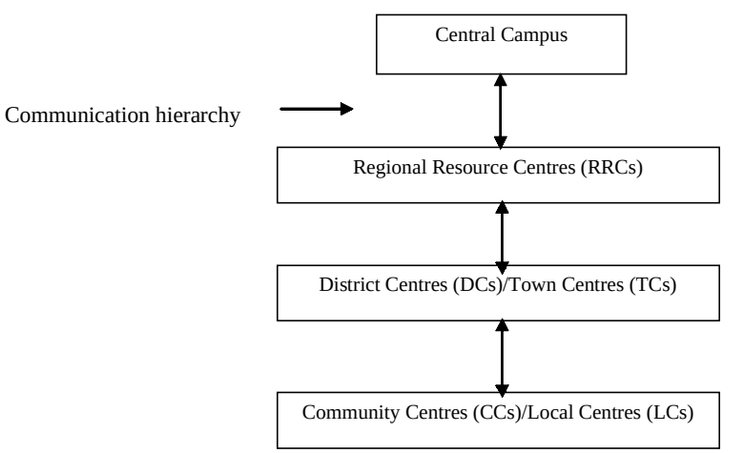
The goal of utilizing multimedia technologies in education is to provide learners with an empowering environment where multimedia may be used anytime, anywhere, at a moderate cost and in an extremely user-friendly manner. However, the technologies employed must remain transparent to the user. Such a computer-based, interactive multimedia environment for distance education is achievable now, but at the cost of high bandwidth infrastructure and sophisticated delivery facilities. Once this has been established for distance education, many other information services essential for accelerated development (e.g., health, governance, business, etc.) may be developed and delivered over the same facilities.

Due to the recent development of information technology, educational courses using a variety of media are being delivered to students in diversified locations to serve the educational needs of the fast-growing populations. Developments in technology allow distance education programs to provide specialized courses to students in remote geographic areas, with increasing interactivity between student and educator. Although the ways in which distance education is implemented differ remarkably from country to country, most distance learning programs rely on technologies that are either already in place or being replicated for their cost effectiveness. Such programs are particularly beneficial for the many people who are not financially, physically or geographically able to obtain conventional education, especially for participants in the developing countries.

Cunningham et al. (2000) referred in their report that "notwithstanding the rapid growth of online delivery among the traditional and new provisions of higher education, there is as yet little evidence of successful, established virtual institutions." However, in a 2002 survey of 75 randomly chosen colleges providing distance learning programs, results revealed an astounding growth rate of 41% per program in the higher education distance learning (Primary Research Group, 2002). Gunawardena and McIsaac (2003), in their *Handbook of Distance Education*, has inferred from the same research case that, "In this time of shrinking budgets, distance learning programs are reporting 41% average annual enrollment growth. Thirty percent of the programs are being developed to meet the needs of professional continuing education for adults. Twenty-four percent of distance students have high-speed bandwidth at home. These developments signal a drastic redirection of traditional distance education." According to an estimate, IT-based education and the e-learning market across the globe is projected at $11.4 billion (United States dollars) in 2003 (Mahajan, Sanone & Gujar, 2003).

It is vital that learners should be able to deal with real-world tasks that require problem-solving skills, integrate knowledge incorporating their own experi-

Figure 1. Communication/management hierarchy of open learning system



ences, and produce new insights in their career. Adult learners and their instructors should be able to handle a number of challenges before actual learning starts; make themselves resourceful by utilizing their own strengths, skills and demands by maintaining self-esteem; and clarify themselves by defining what has been learned, how much it is useful to society and how the content would be effectively utilized for the community in a knowledge-building effort.

One of the barriers to success and development in Open Learning in The Commonwealth developing countries is lack of sound management practice. Sometimes the people who are appointed to high office in open and distance learning do not have proper management skills. As a result, their management practice is poor. They often lack professionalism, proper management ethics and so forth. They lack strategic management skills, they cannot build conducive working environments for staff, nor can they build team spirit required in a learning institution (Tarusikirwa, 2001).

The basic hierarchy of a distance education provider in a country can be shown in Figure 1, adapted from Rahman (2001a).

MAIN FOCUS

There is no mystery to the way effective distance education programs develop. They do not happen spontaneously; they evolve through the hard work and dedicated efforts of many highly committed individuals and organizations. In fact, successful distance education programs rely on the consistent and integrated efforts of learners, faculty, facilitators, support staff and administrators (Suandi, 2001). By adapting available telephone technology, it is easy to implement computer communications through dial-up connectivity. Due to non-availability of high-speed backbone, the bandwidth may be very low, but this technique can be made popular within organizations, academics, researchers, individuals and so forth. The recent global trend of cost reduction in Internet browsing has increased Internet users in many countries. However, as most of the ISPs are located either in the capital or larger metropolitan cities, establishment of regional centres and remote tele-centres located at distant places are now time-demanding.

Teleconferencing, videoconferencing, computer-based interactive multimedia packages and various forms of computer-mediated communications are technologies that facilitate synchronous delivery of content and real-time interaction between teacher and students as well as opportunities for problem-solving, either individually or as a team (Rickards, 2000). Students in developing countries with limited assets may have very little access to these technologies and thus fall further behind in terms of information infrastructure. On the other hand, new telecommunications avenues, such as satellite telephone service, could open channels at a reasonable cost to the remotest areas of the world.

Integrated audio, video and data systems associated with interactive multimedia have been successful distance education media for providing educational opportunities to learners of all ages, at all levels of education and dispersed in diversified geographical locations (Rahman, 2001b). To make the learning processes independent of time and place in combination with technology-based resources, steps need to be taken towards interactive multimedia methods for disseminating education to remote rural-based learners.

Computer technology evolves so quickly that the distant educator focused solely on innovation "not meeting tangible needs" will constantly change equipment in an effort to keep pace with the "latest" technical advancements (Tarusikirwa, 2001). Hence, availability of compatible equipment at a reduced price and integration of them for optimized output becomes extremely difficult during the implementation period, and most of the time, the implementation methodology differs from theoretical design. Sometimes the implementation becomes costly, too, in comparison to the output benefit in the context of a developing country.

Initially, computers with multimedia facilities can be delivered to regional resource centres and media rooms can be established in those centres to be used as multimedia labs. Running those labs would necessitate involvement of two or three IT personnel in each centre. To implement and ascertain the necessity, importance, effectiveness, demand and efficiency, an initial questionnaire can be developed. Distributing periodical surveys among the learners would reflect the effectiveness of the

project for necessary fine-tuning. After complete installation and operation of a few pilot tests in specific regions, the whole country can be brought under a common network through these regional centres.

With a bare minimum information and communications technology (ICT) infrastructure support at the national level, the learning centre can initially focus around 40Km periphery around the main campus, providing line-of-sight radio connectivity ranging from 2Km to 40Km depending on demand and connectivity cost to the nodal/sub-nodal learning centres. These could be schools or community information centres, or affiliated learning centres under the main campus.

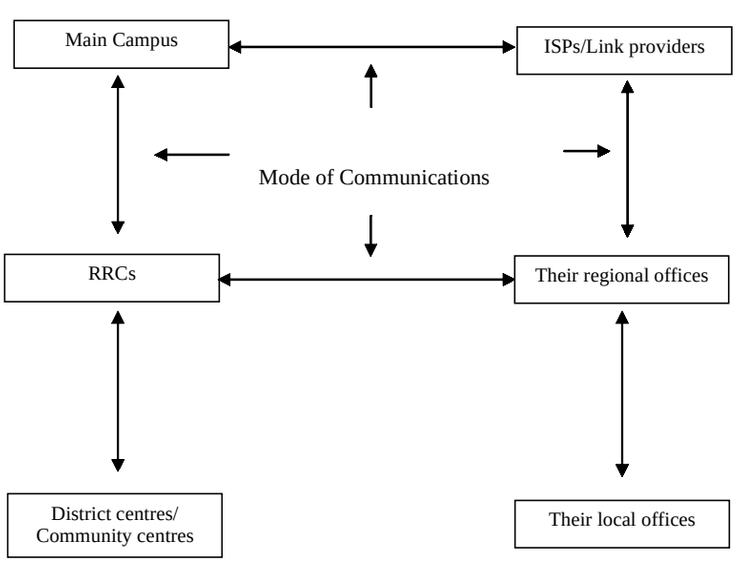
To avail the best opportunity of interactive communications, collaborative approaches could be considered with similar institutions. Offering Internet services at the grass-roots level and effective collaborations among the distance educator and other service providers can set a viable model at the outset. Figure 2, adapted from Rahman (2001a), shows the growth pattern and mode of connectivity between these types of institutions. In the future, more such institutions can easily be brought under this communications umbrella.

A needs-based survey may be necessary during the inception period to enquire about the physical location, demand of the community, requirement of different programs, connectivity issues, the sustainability perspective and other related issues before the establishment of RRCs/DCs/CCs. Following different national consensus, education statistics and demand of local populations, the locations need to be justified (Rahman, 2003). The survey may even become vital for the learning centre authority at a later stage during operation and management.

FUTURE TRENDS

In the absence of a high-speed Internet backbone and basic tele-communications infrastructure, it is extremely difficult to accommodate a transparent communications link with a dial-up connection, and at the same time it is not at all cost effective to enter the Internet with dial-up connectivity. However, in recent days, availability of VSAT (Single Channel Per Carrier/Multiple Channel Per Carrier), radio link (line of sight and non-line of sight) and other Wireless-Fidelity (Wi-Fi) technology has become more

Figure 2. Growth pattern and mode of communications between main campus of the distance education provider and other service providers



receptive to the terminal entrepreneurs and in a way more acceptable to the large group of communities.

Using appropriate techniques, Web-based multimedia technology would be cheaper and more interactive at the front end, accumulating all acquired expenses (Suandi, 2001). Diversified communications methods could easily be adapted to establish a national information backbone. By superimposing it with other available discrete backbones in time without restricting each other's usage, the main backbone can be made more powerful and, hence, be effectively utilized. A combination of media can be used in an integrated way by distant mode course developers. The materials may include specially designed printed self-study texts, study guides and a variety of select articles; or course resource packs for learners containing print, video cassettes, audio cassettes and CDs for each course stage. Computer communication between learners and learners and educators plays a key role in using the education network system (e-mail, Internet, MSN, tele-conferencing, video conferencing, media streaming, etc.).

These distance education strategies may form hybrid combinations of distance and traditional education in the form of distributed learning, networked learning or flexible learning, in which multiple intelligence are addressed through various modes of information retrieval (Gunawardena & McIsaac, 2003). At the same time, infrastructures need to be developed to cope with the increasing number of distant students and availability of low-cost multimedia technologies. In this regard, a dedicated Web server can be treated as an added resource among the server facilities. The Web server is to act as a resource to all students, tutors, staffs and outsiders, providing necessary support in the knowledge dissemination process and a tool for collaborative learning/teaching. Information infrastructure has to be established, so remote stations could log into the Web server and download necessary documents, files and data at reasonably high speed.

CONCLUSION

Effective utilization of capital resources, enhancement towards an improved situation and success of collaborative learning depends largely on socio-economics, geographical pattern, political stability, motivation and ethical issues (Rahman, 2000b). Through sincere effort, concrete ideology, strong positive attitude, dedicated eagerness, sincerity and efficiency, distance educators may achieve the target of enlightening the common citizen of the country by raising the general platform of education. This sort of huge project may involve not only technology issues, but also moral, legal, ethical, social and economic issues, as well. Hence, this type of project may also need to determine the most effective mix of technology in a given learning environment to offer technology-based distant teaching as efficient as traditional face-to-face teaching.

Other diversified facts should be explored, especially by low-income-generating countries, when considering adoption of these advanced technology-based methods in distance education. Socio-economic structure comes first, then availability with affordability, as well as whether those remotely located students could at least be provided with hands-on multimedia technology familiarity. While university academics may debate the educational merits of interactive multimedia environments from theoretical viewpoints, practical issues like accessibility and flexibility of learning experiences have potentially significant impact on the effectiveness of student learning.

With a huge population living in rural areas, spreading education to the rural-based community needs tremendous planning and effort (Rahman et. al., 2000), and a gigantic amount of financing for its successful implementation. Affordability of high-tech infrastructure would necessitate a huge amount of resources, which might not be justified at the initial period, where demand of the livelihood would divert towards some other basic emergency requirements. High initial investment cost would discourage entrepreneurs to be easily convinced, and gear up beyond a pre-conceived state of impression with additional funding.

Absence of a high bandwidth backbone of information infrastructure in developing countries would put the high-tech plan in indisputable difficulties for smooth implementation and operation. A limited number of PCs per student/academic/staff would contradict with the motive of affordable distribution of technology-based methods to remotely located stations.

REFERENCES

Barker, B., & Baker, M. (1995). Strategies to ensure interaction in telecommunicated distance learning. *Paper presented at Teaching Strategies for Distance Learning, 11th Annual Conference on Teaching and Learning*, 17-23.

Cunningham, S. et al. (2000). The Business of Borderless Education, Canberra, Department of Education, 2000.

Gunawardena, C.N., & McIsaac, M.S. (2003). *Handbook of distance education*.

Kochmer, J. (1995). *Internet passport: Northwestnet's guide to our world online.* Bellevue: NorthWestNet and Northwest Academic Computing Consortium.

Mahajan, S., Sanone, A.B., & Gujar, R. (2003). Exploring the application of interactive multimedia in vocational and technical training through open and distance education. *Proceedings of the 17th AAOU Annual Conference,* Bangkok, November 12-14.

Murphy, K. (1995). *Designing online courses mindfully.* Invitational Research Conference in Distance Education. The American Center for the Study of Distance Education.

Primary Research Group. (2002). *The survey of distance and cyber-learning programs in higher education* (*2002 edition*). New York: Primary Research Group.

Rahman, H. (2000a, September 27-30). *A turning point towards the virtuality, the lone distance educator: Compromise or gain.* Paper in the Learning 2000: Reassessing the Virtual University Conference, Virginia Tech.

Rahman, H. (2000b, September 14-17). *Integration of adaptive technologies in building information infrastructure for rural based communities in coastal belt of Bangladesh.* Paper in the First Conference of the Association of Internet Researchers, University of Kansas, Lawrence.

Rahman, H. (2001a, April 1-5). *Replacing tutors with interactive multimedia CD in Bangladesh Open University: A dream or a reality.* Paper in the 20th World Conference on Open Learning and Distance Education, Dusseldorf, Germany.

Rahman, H. (2001b, June 1-3). *Spreading distance education through networked remote information centres.* Paper at the ICIMADE2001, International Conference on Intelligent Multimedia and Distance Education, Fargo, ND.

Rahman, H. (2003). Framework of a technology based distance education university in Bangladesh. *Proceedings of International Workshop on Distributed Internet Infrastructure for Education and research,* BUET, Dhaka, Bangladesh, December 30, 2003-January 2, 2004.

Rahman, M.H., Rahman, S.M., & Alam, M.S. (2000). Interactive multimedia technology for distance education in Bangladesh Open University. *Proceedings of the 15th International Conference on Computers and their Applications (CATA2000),* New Orleans, LA, March 29-31.

Rickards, J. (2000). The virtual campus: Impact on teaching and learning. *Proceedings of the IATUL2000,* Queensland, Australia, July 3-7.

Sherry, L. (1996). Issues in distance learning. *International Journal of Distance Education*, AACE.

Suandi, T. (2001). Institutionalizing support distance learning at Universiti Putra Malaysia. *Proceedings of the Second Pan Commonwealth Forum of Open Learning PCF2*, Durban, South Africa, July 29-August 2.

Tarusikirwa, M.C. (2001). Accessing education in the new millennium: The road to success and development through open and distance learning in the Commonwealth. *Proceedings of the Second Pan Commonwealth Forum of Open Learning PCF2*, Durban, South Africa, July 29-August 2.

Thach, L., & Murphy, K. (1994). Collaboration in distance education: from local to international perspectives. *American Journal of Distance Education, 8*(3), 5-21.

UNESCO. (2001). *Teacher education through distance learning, summary of case studies.* October 2001.

KEY TERMS

Developing Countries: Developing countries are those countries in which the average annual income is low, most of the population is usually engaged in agriculture and the majority live near the subsistence level. In general, developing countries are not highly industrialized, dependent on foreign capital and development aid, whose economies are mostly dependent on agriculture and primary resources, and do not have a strong industrial base. These countries generally have a gross national product below $1,890 per capita (as defined by the World Bank in 1986).

Information and Communications Technology (ICT): ICT is an umbrella term that includes any communication device or application, encompassing: radio, television, cellular phones, computer and network hardware and software, satellite systems and so on, as well as the various services and applications associated with them, such as videoconferencing and distance learning. ICTs are often spoken of in a particular context, such as ICTs in education, health care or libraries.

Interactive Multimedia Techniques: Techniques that a multimedia system uses and in which related items of information are connected and can be presented together. Multimedia can arguably be distinguished from traditional motion pictures or movies both by the scale of the production (multimedia is usually smaller and less expensive) and by the possibility of audience interactivity or involvement (in which case, it is usually called interactive multimedia). Interactive elements can include: voice command, mouse manipulation, text entry, touch screen, video capture of the user or live participation (in live presentations).

Multiple Channel Per Carrier (MCPC): This technology refers to the multiplexing of a number of digital channels (video programs, audio programs and data services) into a common digital bit stream, which are then used to modulate a single carrier that conveys all of the services to the end user.

Single Channel Per Carrier (SCPC): In SCPC systems, each communication signal is individually modulated onto its own carrier, which is used to convey that signal to the end user. It is a type of Frequency Division Multiplexing/Frequency Time Division Multiplexing (FDM/FTDM) transmission where each carrier contains only one communications channel.

Interactive Multimedia Technologies for Distance Education Systems

Hakikur Rahman
SDNP, Bangladesh

INTRODUCTION

Information is typically stored, manipulated, delivered and retrieved using a plethora of existing and emerging technologies. Businesses and organizations must adopt these emerging technologies to remain competitive. However, the evolution and progress of the technology (object orientation, high-speed networking, Internet, etc.) has been so rapid that organizations are constantly facing new challenges in end-user training programs. These new technologies are impacting the whole organization, creating a paradigm shift that in turn enables them to do business in ways never possible before (Chatterjee & Jin, 1997).

Information systems based on hypertext can be extended to include a wide range of data types, resulting in hypermedia, providing a new approach to information access with data storage devices such as magnetic media, video disk and compact disc (CD). Along with alphanumeric data, today's computer systems can handle text, graphics and images, thus bringing audio and video into everyday use.

The Distance Education Task Force (DETF) Report (2000) refers that technology can be classified into non-interactive and time-delayed interactive systems, and interactive distance learning systems. Non-interactive and time-delayed interactive systems include printed materials, correspondence, one-way radio and television broadcasting. Different types of telecommunications technology are available for the delivery of educational programs to single and multiple sites throughout disunited areas and locations.

However, delivering content via the World Wide Web (WWW) has been tormented by unreliability and inconsistency of information transfer, resulting in unacceptable delays and the inability to effectively deliver complex multimedia elements including audio, video and graphics. A CD/Web hybrid, a Web site on a CD, combining the strengths of the CD-ROM and the WWW, can facilitate the delivery of multimedia elements by preserving connectivity, even at constricted bandwidth. Compressing a Web site onto a CD-ROM can reduce the amount of time that students spend interacting with a given technology, and can increase the amount of time they spend learning.

University teaching and learning experiences are being replicated independently of time and place via appropriate technology-mediated learning processes, like the Internet, the Web, CD-ROM and so forth, to increase the educational gains possible by using the Internet while continuing to optimize the integration of other learning media and resources through interactive multimedia communications. Among other conventional interactive teaching methods, Interactive Multimedia Methods (IMMs) seem to be adopted as another mainstream in the path of the distance learning system.

BACKGROUND

F. Hofstetter in his book (*Multimedia Instruction Literacy*) defined "Multimedia Instruction" as "the use of a computer to present and combine text, graphics, audio and video, with links and tools that let the user navigate, interact, create and communicate."

Interactive Multimedia enables the exchange of ideas and thoughts via most appropriate presentation and transmission media. The goal is to provide an empowering environment where multimedia may be used anytime, anywhere, at moderate cost and in a user-friendly manner. Yet the technologies employed must remain apparently transparent to the end user. Interactive distance learning systems can be termed as "live interactive" or "stored interactive," and range from satellite and compressed

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

videoconferencing to stand-alone computer-assisted instruction with two or more participants linked together, but situated in locations that are separated by time and/or place.

Interactive multimedia provides a unique avenue for the communication of engineering concepts. Although most engineering materials today are paper based, more and more educators are examining ways to implement publisher-generated materials or custom, self-developed digital utilities into their curricula (Mohler, 2001). Mohler (2001) also referred that it is vital for engineering educators to continue integrating digital tools into their classrooms, because they provide unique avenues for activating students in learning opportunities and describe engineering content in such a way that is not possible with traditional methods.

The recent media of learning constitutes a new form of virtual learning-communication. It very probably demands an interacting subject that is changed in its self-image. The problem of translation causes a shift of meaning for the contents of knowledge. Questions must be asked: Who and what is communicating there? In which way? And about which specific contents of knowledge? The connection between communication and interaction finally raises the philosophical question of the nature of social relationships of Internet communities, especially with reference to user groups of learning technologies in distance education, generally to the medium in its whole range (Cornet, 2001).

Many people, including educators and learners, enquire among themselves whether distant learners learn as much as those receiving traditional face-to-face instruction. Research indicates that teaching and studying at a distance can be as effective as traditional instruction when the method and technologies used are appropriate to the instructional tasks with intensive learner-to-learner interactions, instructor-to-learner interactions and instructor-to-instructor interactions (Rahman, 2003a). With the convergence of high-speed computing, broadband networking and integrated telecommunication techniques, this new form of interactive multimedia technology has broadened the horizon of distance education systems through diversified innovative methodologies.

MAIN FOCUS

Innovations in the sector of information technology has led educators, scientists, researchers and technocrats to work together for betterment of the communities through effective utilization of available benefits. By far, the learners and educators are among the best beneficiaries at the frontiers of adoptive technologies. Education is no longer a time-bound, schedule-bound or domain-bound learning process. A learner can learn at prolonged pace with enough flexibility in the learning processes, and at the same time, an educator can provide services to the learners through much more flexible media, open to multiple choices.

Using diversified media (local-area network, wide-area network, fiber optics backbone, ISDN, T1, radio link and conventional telephone link), education has been able to reach remotely located learners at faster speed and lesser effort. At the very leading edge of the boomlet in mobile wireless data applications are those that involve sending multimedia data—images, and eventually video—over cellular networks (Blackwell, 2004).

Technology-integrated learning systems can interact with learners both in the mode similar to the conventional instructors and in new modes of information technology through simulations of logical and physical sequences. With fast networks and multimedia instruction-based workstations in distributed classrooms and distributed laboratories, with support from information dense storage media like write-able discs/CDs, structured interactions with multimedia instruction presentations can be delivered across both time and distance.

Several technologies exist within the realm of distance learning and the WWW that can facilitate self-directed, practice-centered learning and meet the challenges of educational delivery to the learner. Several forms of synchronous (real-time) and asynchronous (delayed-time) technology can provide communication between educator and learner that is stimulating and meets the needs of the learner.

The Web is 24 hours a day. Substantial benefits are obtained from using the Web as part of the service strategy (RightNow, 2003). Using the Web format, an essentially infinite number of hyperlinks

may be created, enabling content provided by one member to be linked to relevant information provided by another. Any particular subject is treated as a collection of educational objects, like images, theories, problems, online quizzes and case studies. The Web browser interface lets the individual control how content is displayed, such as opening additional windows to other topics for direct comparison and contrast, or changing text size and placement (Tuthill, 1999).

Interactive and animated educational software combined with text, images and case simulations relevant to basic and advanced learning can be built to serve the learners' community. Utilizing client server technology, Ethernet and LAN/WAN networks can easily span around campus areas and regions. Interactive modules can be created using Macromedia Authorware, Flash, Java applets and other available utilities. They can be migrated to html-based programming, permitting platform independence and widespread availability via WWW. A few technology implications are provided in Table 1 that show the transformation of educational paradigms. Macromedia Director can be used to create interactive materials for use on the WWW in addition to basic html editors.

Some applications of multimedia technologies are:

- analog/digital video
- audio conferencing
- authoring software
- CD-ROMs, drives
- collaborative utility software
- digital signal processors
- hypermedia
- laserdiscs
- e-books
- speech processors, synthesizers
- animation
- video conferencing
- virtual reality
- video capture
- video cams

Table 1. Transformation of educational paradigms

Old Model	**New Model**	**Technology Implications**
Classroom lectures	Individual participation	LAN-connected PCs with access to information
Passive assimilation	Active involvement	Necessitates skill development and simulation knowledge
Emphasize on individual learning	Emphasize on group learning	Benefits from learning tools and application software
Teacher at center and at total control	Teacher as educator and guide	Relies on access to network, servers and utilities
Static content	Dynamic content	Demands networks and publishing tools
Homogeneity in access	Diversity in access	Involves various IMM tools and techniques

Table 2. Types of interaction methods

Interaction methods	**Media**	**Advantage**	**Disadvantage**	**Further development**
Through teachers	E-mail, Usenet, Chat, Conferencing	Quality in teaching	Time consuming	Conferencing Systems, Video processing techniques
Interactive discussions	Interactive Software	Reusability, easier installation	Lengthy development time	High-definition audio and video broadcasts
Collaborative learning	E-mail, Usenet, Chat, Conferencing	Inexpensive, easy access	Less control and supervision	Conferencing systems and discussion tools

Table 3. Delivery methods in interactive learning

Methods	Controlling agents	Media	Advantage/ Disadvantage	Further development
Point to point	Educator or learner	Desktop PC	Better interaction, one-to-one communication /Very expensive	To make it an acceptable solution in a big university or in a developing country situation
Point to multi-point	Teacher or guide	Desktop PC, conferencing system	Flexible/Little interaction	Improved interaction
Multi-point to multi-point	Teacher of guide	Conferencing system, Desktop PC, LAN/WAN	More flexible/Little or no interaction	Improved technology
Streaming, audio, text and video	Student or learner	Internet or intranet	Time and place independent/No Interaction (except simulated techniques)	Improved material presentation

Table 4. Different multicast applications

Topology	Real-time	Non Real-Time
Multimedia	**Video server, Video conferencing, Internet audio, Multimedia events, Web casting (live)**	Replication (Video/Web servers, kiosks), Content delivery (intranets and Internet), Streaming, Web casting (stored)
Data Only	Stock quotes, News feeds, Whiteboards, Interactive gaming	Data delivery (peer/peer, sender/client), Database replication, Software distribution, Dynamic caching

Introducing highly interactive multimedia technology as part of the learning curriculum can offer the best possibilities of development for the future of distance learning. The system should include a conferencing system, a dynamic Web site carrying useful information to use within the course, and access to discussion tools. Workstations are the primary delivery system, but the interaction process can be implemented through various methods as described in Table 2.

Furthermore, course materials used in interactive learning techniques may involve some flexible methods (with little or no interactions) as presented in Table 3.

Miller (1998) and Koyabe (1999) put emphasis on the increased use of multicasting in interactive learning and extensive usage of computers and network equipment in multicasting (routers, switches and high-end LAN equipment). The shaded cell in Table 4 represents real-time multicast applications supported by Real-Time Transport Protocol (RTTP), Real-Time Control Protocol (RTCP) or Real-Time Streaming Protocol (RTSP), while the un-shaded cells show multicast data applications supported by reliable (data) multicast protocols.

Finally, underneath these applications, above the infrastructure, asynchronous transfer mode (ATM) seems to be the most promising emerging technology enabling the development of integrated, interactive multimedia environment for distance education services appropriate for the developing country context. ATM offers economical broadband networking, combining high-quality, real-time video streams with high-speed data packets, even at constricted bandwidth. It also provides flexibility in bandwidth management within the communication protocol, stability in the content, by minimizing data noise, unwanted filter and cheaper delivery by reducing costs of networking.

FUTURE TRENDS

New technologies have established esteemed standing in education and training despite various shortcomings in their performances. Technological innovations have been applied to improve the quality of education for many years. There are instances where applications of the technology had the potential to completely revolutionize the educational systems. Reformed usage of devices like radio, television and video recorders are among many as the starter. Interconnected computers with Internet are the non-concatenated connection between the traditional and innovative techniques. The recent addition of gadgets like personal digital assistants (PDAs), and software like virtual libraries could be some ways out to advanced researchers among many innovative methods on interactive learning.

When prospects of future usage of new technologies emerge in educational settings, there seems to be an innate acknowledgment that positive outcomes will be achieved and these outcomes will justify the expenses. When research is conducted to verify these assumptions, the actual outcomes may sometime be less than those expected. The research methodology behind interactive learning should be based on the notion that the interactivity be provided in the learning context to create environments where information can be shared, critically analyzed and applied, and along the process it becomes knowledge in the mind of the learner.

The use of interactive television as a medium for multimedia-based learning is an application of the technology that needs further investigation by the researchers. Research needs to study the impact of the interactions on the quality of the instructional delivery and develop guidelines for educators and instructional designers to maximize the advantage obtained from this mode of learning in broadcast, narrowcast and multicast modes.

Another emergent technology that appears to hold considerable promise for networked learning is the data broadcasting system (DBS). This technology provides the facility to insert a data stream into a broadcast television signal. Research needs to investigate the utility and efficacy of this technology for use in interactive learning sequences.

Current IMM context has found concrete ground and high potential in distance education methodologies. Further research needs to be carried out towards the cost-effective implementation of this technology. Emphasis should be given to study applications of the technology being used as a vehicle for the delivery of information and instruction and identifying existing problems. Research also needs to focus on developing applications that should make full use of the potentiality offered by this technology.

While security has been extensively addressed in the context of wired networks, the deployment of high-speed wireless data and multimedia communications ushers in new and greater challenges (Bhatkar, 2003). Broadband has emerged as the third wave of technology, offering high bandwidth connectivity across wide-area networks, opening enormous opportunities for information retrieval and interactive learning systems (Rahman, 2003b). However, until the browser software includes built-in support for various audio and video compression schemes, it needs cautious approach from the instructional designer to select the plug-in software that supports multiple platforms and various file formats. Using multimedia files that require proprietary plug-ins usually force the user to install numerous pieces of software in order to access multimedia elements.

It is pertinent that all the newly evolved technologies now exist that are necessary to cost effectively support the revolution in an IMM-based learning system so sorely needed by the developing world. Researchers should take the opportunity to initiate a revolution over the coming years. The main challenges lie in linking and coordinating the "bottom-up" piloting of concepts (at the design stage) with the "top-down" policy-making (at the implementation stage) and budgeting processes from the local (in modular format) to the global level (in repository concept).

CONCLUSION

Regardless of geographical locations, the future learning system cannot be dissociated with information and communication technologies. As technol-

ogy becomes more and more ubiquitous and affordable, virtual learning carries the greatest potential to educate masses in the rural communities in anything and everything. This system of learning can and will revolutionize the education system at the global context, especially in the developing world.

The whole issue of the use of IMM in the learning process is the subject of considerable debate in academic arena. While many educators are embracing applications of multimedia technologies and computer-managed learning, they are advised to be cautious in their expectations and anticipations by their contemporary colleagues. Research in this aspect clearly indicate that media themselves do not influence learning, but it is the instructional design accompanying the media that influences the quality of learning.

The success of the technology in these areas is acknowledged, as is the current move within world-famous universities to embrace a number of the instructional methodologies into their on-campus education system. Much expectation is there for those educators concerned, as well as those wary of assuming that gains will be achieved from these methods and technologies. However, there is a need for appropriate research to support and guide the forms of divergence that have taken place during the last decade in the field of distance education.

One of the long-standing problems in delivering educational content via WWW has been the unpredictability and inconsistency of information transfer via Internet connections. Whether connection to the WWW is established over conventional telephone lines or high-speed LANs/WANs, often, communication is delayed or terminated because of bottlenecks at the server level, congestion in the line of transmission and many unexpected hangouts. Furthermore, the current state of technology does not allow for the optimal delivery of multimedia elements, including audio, video and animation at expected rate. Larger multimedia files require longer download times, which means that students have to wait for a much longer time to deal with these files. Even simple graphics may cause unacceptable delays in congested bandwidth. A CD/Web hybrid, a Web site on a CD, can serve as an acceptable solution in these situations.

REFERENCES

Bhatkar, A. (2003). *Transmission and Computational Energy Modeling for Wireless Video Streaming*, 21.

Blackwell, G. (2004). *Taking advantage of wireless multimedia technology.* January 27.

Chatterjee, S., & Jin, L. (1997). *Broadband residential multimedia systems as a training and learning tool.* Atlanta, GA: Georgia State University.

Cornet, E. (2001, April 1-5). *The future of learning – Learning for the future: Shaping the transition.* The 20th World Conference on Open Learning and Distance Education, Düsseldorf.

Distance Education Task Force. (2000). *Distance Education Task Force Report.* University of Florida.

Koyabe, M.W. (1999). *Large-scale multicast Internet success via satellite: Benefits and challenges in developing countries.* Aberdeen, UK: King's College.

Miller, K. (1998). *Multicasting networking and applications.* Addison-Wesley.

Mohler, J.L. (2001). *Using interactive multimedia technologies to improve student understanding of spatially-dependent engineering concepts.* GraphiCon 2001.

Rahman, H, (2003a). Framework of a technology based distance education university in Bangladesh. *Proceedings of the International Workshop on Distributed Internet Infrastructure for Education and Research (IWIER2003),* Dhaka, Bangladesh, December 30, 2003-January 2, 2004.

Rahman, H. (2003b). Distributed learning sequences for the future generation. *Proceedings of the Closing Gaps in the Digital Divide: Regional Conference on Digital GMS,* Asian Institute of Technology, Bangkok, Thailand, February 26-28.

RightNow Technologies Inc. (2003). *Best practices for the Web-enabled contact center,* 1.

Tuthill, J.M. (1999). *Creation of a network based, interactive multimedia computer assisted instruc-*

tion program for medical student education with migration from a proprietary Apple Macintosh platform to the World Wide Web. University of Vermont College of Medicine.

KEY TERMS

Hypermedia: Hypermedia is a computer-based information retrieval system that enables a user to gain or provide access to texts, audio and video recordings, photographs and computer graphics related to a particular subject.

Integrated Services Digital Network (ISDN): ISDN is a set of CCITT/ITU (Comité Consultatif International Téléphonique et Télégraphique/International Telecommunications Union) standards for digital transmission over ordinary telephone copper wire as well as over other media. ISDN in concept is the integration of both analog or voice data together with digital data over the same network.

Interactive Learning: Interactive learning is defined as the process of exchanging and sharing of knowledge resources conducive to innovation between an innovator, its suppliers and/or its clients. It may start with a resource-based argument, specified by introducing competing and complementary theoretical arguments, such as the complexity and structuring of innovative activities and cross-sectoral technological dynamics.

Interactive Multimedia Method (IMM): It is a multimedia system in which related items of information are connected and can be presented together. This system combines different media for its communication purposes, such as text, graphics, sound and so forth.

Multicast: Multicast is communication between a single sender and multiple receivers on a network. Typical uses include the updating of mobile personnel from a home office and the periodic issuance of online newsletters. Together with anycast and unicast, multicast is one of the packet types in the Internet Protocol Version 6 (IPv6).

Multimedia/Multimedia Technology: Multimedia is more than one concurrent presentation medium (for example, CD-ROM or a Web site). Although still images are a different medium than text, multimedia is typically used to mean the combination of text, sound and/or motion video.

T1: The T1 (or T-1) carrier is the most commonly used digital line in the United States, Canada and Japan. In these countries, it carries 24 pulse code modulation (PCM) signals using time-division multiplexing (TDM) at an overall rate of 1.544 million bits per second (Mbps). In the T-1 system, voice signals are sampled 8,000 times a second and each sample is digitized into an 8-bit word.

International Virtual Offices

Kirk St.Amant
Texas Tech University, USA

INTRODUCTION

Communication technologies are continually expanding our ideas of the office into cyberspace environments. One result of this expansion is the international virtual office (IVO), a setting in which individuals located in different nations use online media to work together on the same project. Different cultural communication expectations, however, can affect the success with which IVO participants exchange information. This article examines three cultural factors that can affect communication within IVO environments.

BACKGROUND

Virtual workplaces offer organizations a variety of benefits, including:

- Increased flexibility and quicker responsiveness (Jordan, 2004)
- Better organizational information sharing (Ruppel & Harrington, 2001)
- Reduced absenteeism (Pinsonneault & Boisvert, 2001)
- Greater efficiency (Jordan, 2004; Salkever, 2003)
- Improved brainstorming practices (Salkever, 2003)

It is perhaps for these reasons that organizations are increasingly using such distributed methods of production (Supporting a Growing, 2004; Pinsonnealut & Boisvert, 2001). The online nature of these workplaces means that they allow for individuals in different nations to participant in certain processes.

This openness is occurring at a time when more of the world is rapidly gaining online access. Taiwan, for example, has the world's fourth highest rate of broadband penetration, while 70% of South Korea and 50% of Hong Kong have broadband access (Global Perspectives, 2004; Taiwan's Broadband, 2004). Such international access, moreover, is expected to grow markedly in the near future. Indian Internet access, for example, is projected to grow by as much as 11 fold in the next four years (Pastore, 2004), and the number of wireless local area networks (WLANs) in China is expected to increase 33% by 2008 (Wireless Networks, 2004). This increased global access brings with it quick and easy connections to relatively inexpensive yet highly skilled technical workforces in other nations (The New Geography, 2003; Weir, 2004). For these reasons, an increasing number of organizations is now examining different ways to use IVOs to tap this international labor force and lower overall production costs (The New Geography, 2003).

To make effective use of such IVO situations, organizations need to understand how cultural factors could affect information exchange among international employees. The problem has to do with differences in cultural communication assumptions. That is, cultural groups can have differing expectations of what constitutes an appropriate or effective method for exchanging information, and these variations even can occur between individuals from the same linguistic background (Driskill, 1996; Weiss, 1998). For example, individuals from different cultures might use alternate strategies for proving an argument (Hofstede, 1997; Weiss, 1998), or cultural groups could have varying expectations of how sentence length (Ulijn & Strother, 1995) or word use (Li & Koole, 1998) contributes to the credibility or intent of a message. These differing expectations, moreover, transcend linguistic boundaries and can affect how individuals interact in a common language (Ulijn, 1996).

While relatively little has been written on how cultural factors could affect IVOs, some research indicates that differing cultural communication expectations can lead to miscommunication or misperception in online exchanges (Artemeva, 1998; Ma, 1996). It is these basic communication issues that organizations must address before they can begin to

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

explore the knowledge management potential that IVOs have to offer. To avoid such problems, employees need to understand how cultural factors could affect online exchanges. They also need to develop strategies to address cultural factors affecting IVO exchanges.

MAIN FOCUS OF THE ARTICLE

Three key areas related to successful communication in IVOs are making contact, status and communication expectations, and the use of a common language. When addressed early and effectively in an IVO, these factors can create the environment essential for effective information exchanges.

Area 1: Making Contact

Successful international online interactions are based on one primary factor—contact. Contact is essential to exchanging information and materials among parties. Making contact requires all parties involved to have similar understandings of how and when exchanges should take place. Yet cultures can have varying expectations of how and when contact should be made. For example, cultural groups can have different expectations of the importance or the exigency associated with a particular medium, a factor that could influence how quickly or how effectively different IVO participants can perform their tasks. Many Americans, for example, believe that an e-mail message merits a quick and timely response. In Ukrainian culture, however, face-to-face communication tends to be valued over other forms of interaction, especially in a business setting (Richmond, 1995). Thus, e-mail to Ukrainian co-workers might not provide as rapid a response as American counterparts might like or require, a factor that could lead to unforeseen delays in an overall process (Mikelonis, 1999). The effects of this delay could be compounded, if others need to wait for this Ukrainian counterpart to complete his or her task before they can begin their own work.

Another factor is the time at which contact can be made. Many Americans, for example, expect to be able to contact co-workers or clients between the hours of 9:00 A.M. and 5:00 P.M. during the standard work week. In France, however, many individuals expect an office to shut down for two or more hours in the middle of the day for the traditional lunch period (generally from noon to 2:00 P.M. or from 1:00 P.M. to 3:00 P.M.) (Weiss, 1998). Such a discrepancy could lead to an unexpected delay in contacting an IVO colleague and in getting essential information quickly.

Similarly, most Americans think of vacations as two- or three- week periods during which someone is in the office to answer the phones. In France, however, it is not uncommon for businesses to close for four to six weeks during the summer, while all of the employees are away on vacation (Weiss, 1998). In these cases, no one may be available to respond to e-mails, receive online materials, or transmit or post needed information.

Additionally, the meaning individuals associate with certain terms can affect information exchanges in IVOs. That is, words such as today, yesterday, and tomorrow can have different meanings, depending on whether they are based on the context of the sender or the recipient of a message. If, for example, a worker in the United States tells a Japanese colleague that he or she needs a report by tomorrow, does the sender mean tomorrow according to the sender's time (in which case, it could be today in Japan), or does the sender mean tomorrow according to Japanese time (in which case, it could be two days from the time at which the message was sent)?

To avoid such contact-related problems, individuals working in IVOs can adopt a series of strategies for interacting with international colleagues:

- **Agree upon the medium that will serve as the primary mechanism for exchanging information and establish expectations for when responses to urgent messages can be sent.** Individuals need to agree upon the best means and medium of contacting others when a quick response is essential and then set guidelines for when one can expect an international colleague to check his or her messages and when/how quickly a response can be sent, based on factors of culture and time difference.
- **Establish a secondary medium for making contact, should the primary medium fail.** Certain circumstances could render a medium inoperative. For this reason, individuals should establish a backup method for contacting over-

seas colleagues. In Ukraine, for example, what should individuals do if the primary method for making contact is e-mail, but a blackout unexpectedly happens at a critical production time (not an uncommon occurrence in many Eastern European countries)? The solution would be to establish an agreed-upon secondary source that both parties can access easily (e.g., cell phones).

- **Establish a context for conveying chronological references.** IVO participants should never use relative date references (i.e., tomorrow or yesterday), but instead should provide the day and the date (e.g., Monday, October 4), as well as some additional chronological context according to the recipient's time frame (e.g., Netherlands time). For example, tell a Dutch colleague that information is needed by Monday, October 4, 16:00 Netherlands time.

By following these steps, employees in IVOs can increase the chances of making contact with overseas co-workers and receiving timely responses.

Area 2: Status and Communication Expectations

In some cultures, there is the flexibility to circumvent official channels in order to achieve a particular goal. In the United States, for example, a person with a good idea might be able to present that idea directly to his or her division manager instead of having to route that idea through his or her immediate supervisor. In other cultures, however, structures are more rigid, and employees must go through a set of expected formal channels if they wish to see results. In such systems, attempts to go around a hierarchy to achieve an end could damage the reputation of or threaten the job of employees using such methods. Hofstede (1997) dubbed this notion of how adamantly different cultures adhered to a hierarchical system of status and formality as *power distance.* In general, the higher the degree of power distance, the less permissible it is for subordinates to interact with superiors, and the greater the degree of formality expected if such parties should interact.

IVOs, however, can create situations that conflict with such systems, for online media remove many of the cues that individuals associate with status and can contribute to the use of a more information tone in online exchanges (St.Amant, 2002). Individuals working in IVOs should adopt, therefore, certain communication practices that address factors of culture and status:

- **Learn the hierarchical structure of the cultural groups with which one will interact.** Once individuals identify these systems, they should learn how closely members of that culture are expected to follow status roles. Additionally, cultures might have different expectations of if and how such structures can be bypassed (e.g., emergency situations). By learning these status expectations, IVO participants can determine how quickly they can get a response to certain requests.
- **Determine who one's status counterparts are in other cultures.** Such a determination is often needed to ensure that messages get sent to the correct individual and not to someone at a higher point in the power structure. IVO participants also should restrict contact with high status persons from other cultures until told otherwise by high status members of that culture.
- **Avoid given or first names when addressing someone from another culture**. In cultures where status is important, the use of titles is also expected (Hofstede, 1997). For this reason, IVO participants should use titles such as Mr. or Ms. when addressing international counterparts. If the individual has a professional title (e.g., Dr.), use that title when addressing the related individual. One should continue to use such titles until explicitly told otherwise by an international counterpart.

By addressing factors of status, IVO participants can keep channels of cross-cultural communication open and maintain contact.

Area 3: Using a Common Language to Communicate

IVOs often require individuals from different linguistic backgrounds to use a common tongue when interacting within the same virtual space. Such situations bring with them potential problems related to fluency in that language. That is, the fact that an

individual speaks a particular language does not necessarily mean that that person speaks it well or understands all of the subtle nuances and intricate uses of the language (Varner & Beamer, 1995). Even within language groups, dialect differences (e.g., British vs. American English or Luso vs. Iberian Portuguese) could cause communication problems.

In IVOs, the issue of linguistic proficiency is further complicated by the nature of online media, which remove accents that are often indicators of another's linguistic abilities (e.g., being a non-native speaker of a language). Additionally, communication expectations associated with different online media might skew perceptions of an individual's linguistic proficiency. E-mails, for example, are often quite brief, and individuals tend to be more tolerant of spelling and grammar errors in e-mails than more conventional printed messages (And Now, 2001). As a result of such factors, IVO participants might either forget that an international counterpart is not a native speaker of a language or not realize that an individual does not speak that language as well as one might think.

To avoid language problems in IVOs, individuals should remember the following:

- **Avoid idiomatic expressions.** Idiomatic expressions are word combinations that have a specific cultural meaning that differs from their literal meaning. For example, the American English expression "It's raining cats and dogs" is not used to mean that cats and dogs are falling from the sky (literal meaning); rather, it means "it is raining forcefully" (intended meaning). Because the intended meaning of such an expression is based on a specific cultural association, individuals who are not a part of a particular culture can be confused by such phrases (Jones, 1996).
- **Avoid abbreviations.** Abbreviations are like idioms; they require a particular cultural background to understand what overall expression they represent (Jones, 1996). If abbreviations are essential to exchanging information, then individuals should spell out the complete term the first time the abbreviation is used and employ some special indicator to demonstrate how the abbreviation is related to the original expression (e.g., "This passage examines the role of the *I*nternal *R*evenue *S*ervice (IRS)").
- **Establish what dialect of a common language will be used by all participants.** Certain dialect differences sometimes can result in confusion within the same language. For example, speakers of various dialects of a language could have different terms for the same object or concept, or could associate varying meanings with the same term. By establishing a standard dialect for IVO exchanges, individuals can reduce some of the confusion related to these differences.

Such strategies can reduce confusion related to linguistic proficiency or dialect differences.

While the ideas presented in this section are quite simple, they can be essential to communicating across cultural barriers. The efficiency with which individuals interact in IVOs, moreover, will grow in importance, as organizations increasingly look for ways to tap into different overseas markets.

FUTURE TRENDS

The global spread of online communication technologies is providing access to new and relatively untapped overseas markets with consumers who are increasingly purchasing imported goods. For example, while Chinese wages remain relatively low, there is a small yet rapidly growing middle class that is becoming an important consumer base for technology products (China's Economic Power, 2001; Hamm, 2004). In fact, China's import of high-tech goods from the U.S. alone has risen from $970 million USD in 1992 to almost $4.6 billion USD in 2000 (Clifford & Roberts, 2001).

Similarly, the Indian boom in outsourcing services has led to a growing middle class with an aggregate purchasing power of some $420 billion USD (Malik, 2004). Additionally, as more work is outsourced to employees in the developing world, more money will flow into those nations, and this influx of capital brings with it the potential to purchase more products (Hamm, 2004). Moreover, since much of this outsourcing work is facilitated by the Internet and the World Wide Web, these outsource workers

become prospective consumers who are already connected to and familiar with online media that can serve as marketing channels.

Within this business framework, IVOs could be highly important for a number of reasons. First, they could provide project groups with direct access to international markets by including a member of a particular culture in an IVO. This individual could then supply his or her counterparts with country-specific information used to modify the product to meet the expectations of a particular group of consumers. Second, these individuals could trial run products in a related culture and make recommendations for how items should be modified to meet consumer expectations. Finally, this individual could also act as an in-country distribution point for getting completed electronic materials (e.g., software) into that market quickly. As a result, the adoption of IVOs will likely increase both in use and in international scope, and today's workers need to understand and address cultural factors so that they can communicate effectively within such environments.

CONCLUSION

Today, the widespread use of e-mail and corporate intranets has begun to change the concept of "the office" from a physical location to a state of mind. This article examined some of the more problematic cross-cultural communication areas related to international virtual offices (IVOs) and provided strategies for communicating efficiently within such organizations. By addressing such factors early on, organizations can enhance the production capabilities of such IVOs.

REFERENCES

And Now for Some Bad Grammar. (2001). Manage the ecommerce business. Retrieved January 22, 2004, from *http://ecommerce.internet.com/how/biz/print/0,,10365_764531,00.html*

Artemeva, N. (1998). The writing consultant as cultural interpreter: Bridging cultural perspectives on the genre of the periodic engineering report. *Technical Communication Quarterly, 7,* 285-299.

China's Economic Power. (2001). *The Economist,* 23-25.

Clifford, M., & Roberts, D. (2001). China: Coping with its new power. *BusinessWeek,* 28-34.

Driskill, L. (1996). Collaborating across national and cultural borders. In D.C. Andrews (Ed.), *International Dimensions of Technical Communication* (pp. 23-44). Arlington, VA: Society for Technical Communication.

Global Perspectives on US Broadband Adoption. (2004). *eMarketer*. Retrieved September 15, 2004, from *http://emarketer.com/Article.aspx?1003041&printerFriendly=yes*

Hamm, S. (2004). Tech's future. *BusinessWeek*, 82-89.

Hofstede, G. (1997). *Cultures and organizations: Software of the mind*. New York: McGraw Hill.

Jones, A.R. (1996). Tips on preparing documents for translation. *GlobalTalk: Newsletter for the International Technical Communication SIG*, 682, 693.

Jordan, J. (2004). Managing "virtual" people. *BusinessWeek online*. Retrieved September 20, 2004, from *http://www.businessweek.com/print/smallbiz/content/apr2004/sb20040416_7411_sb008.html*

Li, X., & Koole, T. (1998). Cultural keywords in Chinese-Dutch business negotiations. In S. Niemeier, C.P. Campbell, & R. Dirven (Eds.), *The cultural context in business communication* (pp. 186-213). Philadelphia, PA: John Benjamins.

Ma, R. (1960). Computer-mediated conversations as a new dimension of intercultural communication between East Asian and North American college students. In S. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 173-186). Amsterdam: John Benjamins.

Malik, R. (2004, July). The new land of opportunity. *Business 2.0*, 72-79.

Mikelonis, V.M. (1999, June 27). Eastern European question. Personal e-mail.<mikel001@maroon.tc.umn.edu>

The new geography of the IT industry. (2003, July 17). *The Economist*. Retrieved August 10, 2003,

from *http://www.economist.com/displaystory.cfm?story_id=1925828*

Pastore, M. (2004, February 24). India may threaten China for king of netizens. *ClickZ Stats.* Retrieved April 10, 2004, from *http://www.clickz.com/stats/big_picture/geographics/article.php/309751*

Pinsonneault, A., & Boisvert, M. (2001). The impacts of telecommuting on organizations and individuals: A review of the literature. In N.J. Johnson (Ed.), *Telecommuting and virtual offices: Issues & opportunities* (pp. 163-185). Hershey, PA: Idea Group.

Richmond, Y. (1995). *From da to yes: Understanding the East Europeans*. Yarmouth, ME: Intercultural Press.

Ruppel, C.P., & Harrington, S.J. (2001). Sharing knowledge through intranets: A study of organizational culture and intranet implementation. *IEEE Transactions on Professional Communication, 44*, 37-52.

Salkever, A. (2003, April 24). Home truths about meetings. *BusinessWeek online*. Retrieved September 20, 2004, from *http://www.businessweek.com/print/smallbiz/content/apr2003/sb20030424_0977_sb010.html*

St. Amant, K. (2002). When cultures and computers collide. *Journal of Business and Technical Communication, 16*, 196-214.

Supporting a growing mobile workforce. (2004, September 9). *eMarketer*. Retrieved September 15, 2004, from *http://www.emarketer.com/Article.aspx?1003033&printerFriendly=yes*

Taiwan's broadband penetration rate fourth highest worldwide. (2004, September 9). *eMarketer*. Retrieved September 15, 2004, from *http://www.emarketer.com/Article.aspx?1003032&printerFriendly=yes*

Ulijn, J.M. (1996). Translating the culture of technical documents: Some experimental evidence. In D.C. Andrews (Ed.), *International Dimensions of Technical Communication* (pp. 69-86). Arlington, VA: Society for Technical Communication.

Ulijn, J.M., & Strother, J.B. (1995). *Communicating in business and technology: From psycholinguistic theory to international practice*. Frankfurt, Germany: Peter Lang.

Varner, I., & Beamer, L. (1995). *Intercultural communication in the global workplace*. Boston: Irwin.

Weir, L. (2004, August 24). Boring game? Outsource it. *Wired News*. Retrieved September 20, 2004, from *http://www.wired.com/news/print/0,1294,64638,00.html*

Weiss, S.E. (1998). Negotiating with foreign business persons: An introduction for Americans with propositions on six cultures. In S. Niemeier, C.P. Campbell, & R. Dirven (Eds.), *The cultural context in business communication* (pp. 51-118). Philadelphia, PA: John Benjamins.

Wireless networks to ride China's boom. (2004, September 16). *eMarketer*. Retrieved September 17, 2004, from *http://www.emarketer.com/Article.aspx?1003043&printerFriendly=yes*

KEY TERMS

Access: The ability to find or to exchange information via online media.

Contact: The ability to exchange information directly with another individual.

Dialect: A variation of a language.

Idiomatic Expression: A phrase that is associated with a particular, non-literal meaning.

International Virtual Office (IVO): A work group comprised of individuals who are situated in different nations and who use online media to collaborate on the same project.

Online: Related to or involving the use of the Internet or the World Wide Web.

Power Distance: A measure of the importance status has in governing interactions among individuals.

Internet Adoption by Small Firms

Paul B. Cragg
University of Canterbury, New Zealand

Annette M. Mills
University of Canterbury, New Zealand

INTRODUCTION

Research shows that small firms make significant contributions to their economic environment. With the significant advances being made in Information and Communication Technologies (ICTs), the Internet has become very important to many small firms, enabling them to overcome various inadequacies attributed to factors such as firm size, availability of resources and other technological, operational and managerial shortfalls. Despite the contributions that the adoption of Internet technology can make to the well being of such firms, research shows that many small firms have not yet embraced the technology in ways that will allow them to capitalise on potential benefits. It is therefore important for firms and researchers to understand the factors that enable (or hinder) the adoption of various technologies.

BACKGROUND

The adoption of a technology (in this case, the Internet) can be viewed as an *innovation* for a firm, where that technology represents something that is new to the adopting organization (Damanpour, 1991). Thus, adopting the Internet for e-mail could be seen as an innovation, so too Web browsing and engaging in electronic commerce (e-commerce) to sell or purchase goods. Innovation theory suggests that the adoption of an innovation may have a number of stages. For example, Zaltman, Duncan and Holbek (1973) suggested the adoption of an innovation may take place in two stages: the *initiation* stage, involving knowledge and awareness of the innovation, the formation of attitudes toward the innovation and decision making (i.e., whether to adopt the innovation); this is followed by the *implementation* stage, when the actual implementation of the technology takes place. Rogers (2003) proposed a similar view of adoption; namely, the *innovation-decision* process, which comprises the stages of *knowledge*, *persuasion*, *decision*, *implementation* and *confirmation*. It is at the *decision* stage that the organization determines whether to accept or reject the innovation.

An adoption can also be examined in terms of the ways in which the technology has been used. This view is especially relevant to Internet adoption, which can include simpler forms such as e-mail adoption and Web searching (without a Web site presence); or a firm's Internet presence, whether the firm has a Web site that provides general information only or information pertinent to customers; or one in which Internet activity is an integral part of the firm's business processes (e.g., Teo & Pian, 2003).

INTERNET ADOPTION

Whether one views the adoption of an innovation in terms of stages or the way in which a technology is used, such adoption is influenced (i.e., enabled or inhibited) by various classes of factors, *innovation (technological) factors* (e.g., perceived benefits, complexity and compatibility, including business strategy), *organizational factors* (e.g., firm size, technological readiness, IT support, management support, financial readiness) and *environmental factors* (e.g., pressure from clients, competitors and trading partners). Similar frameworks have been successfully used to identify factors that influence the adoption of various ICTs by small firms, including electronic data interchange (EDI) (Chwelos, Benbasat & Dexter, 2001; Iacovou, Benbasat & Dexter, 1995), the Internet (Mehrtens, Cragg & Mills, 2001; Poon & Swatman, 1999; Teo & Pian, 2003; Walczuch, Van Braven & Lundgren, 2000),

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

e-commerce (Pearson & Grandon, 2004; Kendall, Tung, Chua, Ng & Tan, 2001; Raymond, 2001) and other ICTs (Thong, 1999). The following sections discuss these influences in more detail.

Innovation (Technological) Factors

Innovation factors include *perceived benefits* and *compatibility* (McGowan & Madey, 1998). *Perceived benefits* refers to the direct (e.g., operational savings related to internal efficiency of the organization) and indirect benefits (opportunities derived from the impact of the Internet on business processes and relationships) that a technology can provide the firm (Iacovou et al., 1995). Research shows that small firms expect to derive various benefits, such as improved communications, cost savings, time savings and increased market potential from Internet adoption, direct and indirect advertising, and internationalization (Chwelos et al., 2001; Iacovou et al., 1995; Kendall et al., 2001; Mehrtens et al., 2001; Poon & Swatman, 1997; Pollard, 2003; Walczuch et al., 2000).

For example, Mehrtens et al. (2001) identified the *relative advantage* of the Internet as a communication and business tool when compared to traditional methods of communication (such as telephones and faxes) as a key decision factor. The opportunity to present information on a Web site was also seen as an advantage over traditional forms of advertising and retailing. The concept of global sourcing of information also forms part of the relative advantage of the Internet.

On the other hand, concern that expected benefits (such as lower costs or greater efficiency) would not be achieved, mismatch between business strategy and Internet technology, and lack of direct benefits were factors that inhibit adoption among small firms (Chan & Mills, 2002; Cragg & King, 1993; Walczuch et al., 2000).

Compatibility describes the degree to which an innovation is perceived as consistent with the existing values, past experiences and needs of a potential adopter (Rogers, 2003). For example, research shows that compatibility with existing systems is positively associated with technology adoption (e.g., Duxbury & Corbett, 1996). Compatibility also includes the extent to which a technology aligns with the firm's needs, including the alignment of a firm's IT strategy with its business strategy (King & Teo, 1996; Walczuch et al., 2000). For example, research has shown that business strategy directly influences the adoption and integration of IT into the organization (Teo & Pian, 2003), and that without a corporate e-commerce strategy for guidance, firms may adopt non-integrated information systems with conflicting goals (Raghunathan & Madey, 1999).

Teo and Pian (2003) identified alignment with business strategy as the most important factor impacting the level of Internet adoption. For example, Walczuch et al. (2000) found that small firms were reluctant to adopt particular Internet technologies (e.g., Web site) where the firm believed that the technology was not compatible with its business purpose. Similarly, Chan and Mills (2002) found that small brokerages were reluctant to adopt online (Internet-enabled) trading where this was not regarded as compatible with business strategy. Firms that believed Internet-enabled technologies yielded *benefits* and were *compatible* with the firm's values and needs were found to be earlier adopters of the technology, while firms that were not convinced regarding benefits and compatibility aspects of the technology tended to be later adopters, or reject the technology or perceive these factors as key inhibitors of adoption (Chan & Mills, 2002; Walczuch et al., 2000). Similarly, Pearson and Grandon (2004) found that compatibility was a key factor distinguishing adopters from non-adopters.

Organizational Factors

Organizational factors address the resources that an organization has available to support the adoption (McGowan & Madey, 1998). These include firm size, financial and technological resources (including IT support), and top management support. While some studies have focused on individual factors, some recent research has emphasized all of these organizational factors under the title of "dynamic capabilities" (Helfat & Raubitschek, 2000). Dynamic capabilities are firm-level attributes that enable firms to be innovative, for example, by introducing new products and processes and adapting to changing market conditions. Adopting new technologies like the Internet is one example of innovative activity. Firms with superior dynamic capabilities are better able to introduce and assimilate new technologies. Both Daniel and

Wilson (2003) and Wheeler (2002) have applied dynamic capability perspectives to examining e-business in large firms. For example, Wheeler (2002) argues that specific capabilities, such as choosing new IT and matching opportunities to IT, have the potential to distinguish successful firms from less successful firms. Daniel and Wilson (2003) identified eight capabilities that distinguish successful and unsuccessful firms, including the ability to integrate new IT systems. The concept of dynamic capability, therefore, has the potential to be useful for understanding small firm adoption.

Firm size is also a key adoption factor; prior research suggests that smaller firms may be less likely to adopt e-commerce (Teo & Pian, 2003). For small firms that adopt information technologies, prior research identifies individual factors (within each of the adoption factor classes) as key drivers of Internet-enabled technology adoption research. These include alignment with business strategy, perceived benefits, relative advantage, compatibility, complexity, trialability, proactivity towards technology, organizational support, financial readiness, technological readiness, IT knowledge of non-IS professionals, management support, CEO attitudes, internal and external IS support and external pressure (Chan & Mills, 2002; Chwelos et al., 2001; Cragg & King, 1993; Pearson & Grandon, 2004; Iacovou et al., 1995; Kendall et al., 2001; Mehrtens et al., 2001; Thong, 1999).

Firm size not only impacts a firm's ability to adopt Internet technology but also the level at which a firm is likely to adopt the technology. For example, Teo and Pian (2004) found that while larger firms tend to adopt Web technology at higher levels, small firms tend to adopt the technology at the lower levels of e-mail, establishing a Web presence and providing limited Web services, such as information provision and some product access to customers.

Financial support refers to financial assistance from within or outside the firm's resources (e.g., loans and subsidies) that equip the firm to acquire Internet technology. Access to sufficient financial resources is generally acknowledged as a factor that enables adoption. For example, only firms that have adequate financial resources are likely to adopt IT (Pearson & Grandon, 2004; Iacovou et al., 1995; Thong, 1999). Although the cost of technology adoption can vary widely, limited financial resources can inhibit uptake, especially for small firms. For example, research shows that the cost of development and maintenance, concerns over expected benefits (such as lower costs or greater efficiency) and inadequate financial resources are factors that may slow or inhibit adoption among small firms (Cragg & King, 1993; King & Teo, 1996; Walczuch et al., 2000). Although, Mehrtens et al. (2001) found no significant relationship between adoption and financial support, this was likely because firms could readily afford the cost of adopting the Internet at a basic level. Similarly, Chan and Mills (2002) explored a more costly adoption (i.e., online stock trading), but found insufficient evidence to conclude whether financial readiness was a key factor. Nonetheless, it is reasonable to expect that having access to adequate financing is a critical step in the adoption process and in determining the level of adoption.

Technological readiness includes internal IT sophistication and access to external IT support. *Internal IT sophistication* refers to the level of sophistication of IT usage, IT management and IT skill within the organization (Iacovou et al., 1995). For example, research has found that firms that are more IT sophisticated (e.g., have a formally established IT department and other IT assets, such as IT knowledge and IT capabilities) are more likely to adopt technologies such as EDI and e-commerce technology (Iacovou et al., 1995; Lertwongsatien & Wongpinunwatana, 2003). CEO knowledge of IT has also been identified as a key factor influencing adoption and the championing of IT (Bassellier, Benbasat & Reich, 2003; Thong, 1999), while lack of knowledge appears to inhibit uptake (e.g., AC Nielson, 2001).

External IT support refers to IT-related assistance received from outside the firm (e.g., external consultants). Since small firms in particular often lack access to sufficient internal IT resources, external support is a key enabler of technology adoption (e.g., Cragg & King, 1993; Pollard, 2003; Raymond & Bergeron, 1996). For example, research suggests that strong support from external technical sources may lead to or accelerate Internet adoption (Chan & Mills; 2002). A study of Internet non-adopters also showed that lack of IT expertise, lack of employee IT knowledge and skills, lack of business relevance and concern that staff would waste time surfing were reasons for not adopting the Internet (Teo & Tan, 1998).

Top management characteristics and support. Researchers argue that top management characteristics and top management support of an innovation could lead to adoption or early adoption (King & Teo, 1996; Raymond & Bergeron, 1996). For example, top management characteristics such as CEO knowledge of IT, CEO values and CEO attitude towards an innovation are considered important factors influencing IT adoption (Bassellier et al., 2003; Thong, 1999). The support of a top management champion can have a positive impact on adoption (e.g., Mehrtens et al., 2001). For example, Chan and Mills (2002) found that the strong commitment of top management led to early adoption, while lack of top management commitment inhibited adoption. On the other hand, while research suggests that top management support is a significant determinant of a firm's decision to adopt a technology, the non-significant findings regarding its influence on the level of adoption may suggest that top management support does not directly influence the level of adoption (Teo & Pian, 2003; Thong, 1999).

Environmental Factors

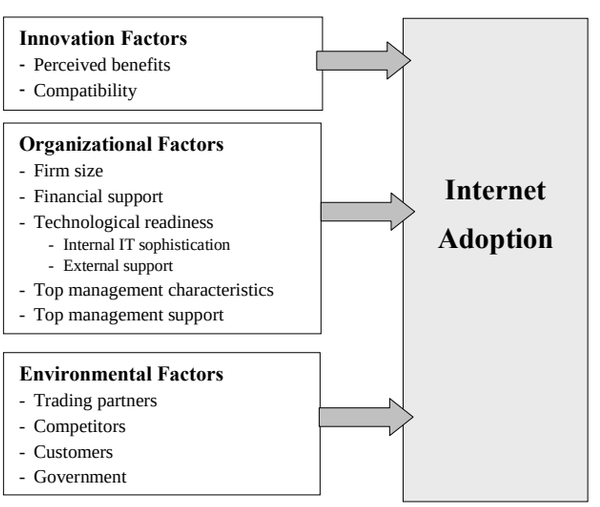
The environmental context includes external pressures and support for technology adoption. For example, research shows that external pressures most often derive from competitors, clients and trading partners (including suppliers and contractors), and other characteristics of the marketplace such as legal requirements (Iacovou et al., 1995). For example, the e-commerce adoption decision of traditional firms may be influenced by other "e-commerce-able" firms (Chircu & Kauffman, 2000). Similarly, small firms with close and significant trading relationships with EDI initiators may feel pressured to adopt EDI in order to maintain their business relationships, even to the extent of adopting the EDI vendor recommended by their trading partner without further investigation (Chen & Williams, 1998). Chwelos et al. (2001) also found that competitive pressure was the single most important factor contributing to EDI adoption.

Raymond (2001) suggested the ways in which small firms used Internet-based technologies were determined by the environment in which these firms operated. In a study of small firms in the travel industry, Raymond found that environmental pressures derived from a need to imitate competitors, coercive pressure from suppliers and business partners, and the expectations of a sales or promotional presence or a Web site presence for current and potential customers.

Mehrtens et al. (2001) also found that the pressure to adopt the Internet came from other Internet users, typically from customers or potential customers; this was expressed more as an expectation that the organization have an e-mail address and a Web site rather than as a specific pressure factor. There was also an expectation that the organization be active on the Internet, including regular browsing and being as up to date as clients. Contrary to other research (e.g., Chwelos et al., 2001), the firms studied by Mehrtens et al. (2001) did not indicate that their adoption of the Internet was influenced by competitors. However, as the Internet gains in popularity, this pressure to adopt could be felt by small firms who are slow to adopt the Internet. Similarly, Rogers' (1991) diffusion study suggested that interactive technology had zero utility until other individuals had adopted the technology as well, so until a *critical mass* of adopters was achieved, the rate of adoption would be slow.

Government initiatives and support as well as support from non-competitive industry players may also encourage adoption (e.g., Pollard, 2003; Scupola, 2003). For example, Scupola (2003) found that

Figure 1. Factors affecting Internet adoption



government interventions by way of subsidies, state support, financial incentives (e.g., tax breaks) and training encouraged e-service usage. Changes in public administration operations (e.g., using the Internet to provide citizen and company information or administer tax systems) were also found to encourage Internet adoption.

FUTURE TRENDS

Mehrtens et al. (2001) found that even small firms in the IT sector lacked sufficient knowledge to rapidly adopt the Internet. As such, the greater mass of small firms without such IT awareness can be expected to take even longer to adopt various Internet technologies. This suggests there are business opportunities to assist small firms to adopt Internet technologies. For example, Lockett and Brown (2000) indicate the potential role for firms to act as intermediaries to assist the formation of "eClusters," a relatively new business model enabled by the Internet where one or more intermediaries do much of the computing for a group of related small firms. Such intermediaries could address numerous IT management tasks (e.g., the determination of needs, selection implementation and operation of hardware and software), enabling firms to concentrate on core business activities rather than carry out IT management themselves.

There are also opportunities for further research into Internet adoption by small firms. In particular, while most studies focus on adopters, only a few studies include non-adopters (Chan & Mills, 2002; Mehrtens et al., 2001; Teo & Tan, 1998) or address the decision not to adopt the Internet. Although some research has investigated post-adoption satisfaction (Liu & Khalifa, 2003), little is known about Internet post-adoption stages, including implementation. In-depth longitudinal case studies of individual firms could also add significantly to current understanding.

There are also opportunities to examine each type of influence in-depth by focusing on individual factors. For example, research indicates that business strategy is a key determinant of Internet adoption. However, many small firms do not have a particular Internet strategy, despite their expectation of particular benefits, such as advertising, marketing, enabling customer feedback and globalization (Webb & Sayer, 1998). A study of organizational readiness could also help determine how a firm attains the desired level of IT use and IT knowledge for Internet adoption.

There is also a need to extend small firm research to include the development of a predictive adoption model. Furthermore, most studies of small firm Internet adoption have focused on relatively simple applications of the Internet (e.g., e-mail, Web browsing and Web site presence). There has been little study of more sophisticated applications of business-to-business (B2B) and business-to-consumer (B2C) e-commerce involving transactions over the Internet (Brown & Lockett, 2004). It is likely that different factors (such as firm size, IT expertise and financial readiness) may have a greater influence on the adoption of these more sophisticated applications.

CONCLUSION

Research has shown that innovation factors, organizational factors and environmental factors are significant factors influencing small firms' decision to adopt the Internet. More specifically, such research identifies perceived benefits (including relative advantage), compatibility, technological readiness, firm size, top management characteristics and support, and the influence of customers, trading partners, competitors and government as factors influencing the small firm adoption decision. However, as many small firms have not adopted sophisticated Internet technologies, a major research opportunity exists to improve our understanding of how small firms can successfully become more sophisticated users of the Internet.

REFERENCES

AC Nielsen. (2001). Electronic commerce in New Zealand: A survey of elec tronic traders. *A report prepared for Inland Revenue Department and Ministry of Economic Development.* Ref# 1402282. Retrieved January, 2004, from *www.ecommerce.govt.nz/statistics/index.html#survey*

Bassellier, G., Benbasat, I., & Reich, B.H. (2003). The influence of business managers IT competence on championing IT. *Information Systems Research*, *14*(4), 317-336.

Brown, D.H., & Lockett, N. (2004). Potential of critical e-applications for engaging SMEs in e-business: a provider perspective. *European Journal of Information Systems*, (1), 21-34.

Chan, Patrick Y.P., & Mills, A.M. (2002). Motivators and inhibitors of e-commerce technology adoption: Online stock trading by small brokerage firms in New Zealand. *Journal of Information Technology Cases and Applications*, (3), 38-56.

Chen, J.C., & Williams, B.C. (1998). The impact of electronic data interchange (EDI) on SMEs: Summary of eight British case studies. *Journal of Small Business Management*, (4), 68-72.

Chircu, A.M., & Kauffman, R.J. (2000). Re-intermediation strategies in business-to-business electronic commerce. *International Journal of Electronic Commerce*, (4), 7-42.

Chwelos, P., Benbasat, I., & Dexter, A.S. (2001). Research report: empirical test of an EDI adoption model. *Information Systems Research*, (3), 304-321.

Cragg, P.B., & King, M. (1993). Small-firm computing: motivators and inhibitors. *MIS Quarterly,* (1), 47-60.

Damanpour, F. (1991). Organizational innovation: A meta-analysis of effects of determinants and moderators. *Academy of Management Journal*, (3), 555-590.

Daniel, E.M., & Wilson, H.N. (2003). The role of dynamic capabilities in e-business transformation. *European Journal of Information Systems*, (4), December, 282-296.

Duxbury, L., & Corbett, N. (1996). Adoption of portable offices: An exploratory analysis. *Journal of Organizational Computing and Electronic Commerce*, (4), 345-363.

Helfat, C.E., & Raubitschek, R.S. (2000). Product sequencing: Co-evolution of knowledge, capabilities and products. *Strategic Management Journal,* (10/11), 961-979;

Iacovou, C.L., Benbasat I., & Dexter, A. (1995). Electronic data interchange and small organizations: Adoption and impact of technology. *MIS Quarterly*, (4), December, 466-485.

Kendall, J.D., Tung, L.L., Chua, K.H., Ng, C.H.D., & Tan, S.M. (2001). Receptivity of Singapore's SMEs to electronic commerce adoption. *Journal of Strategic Information Systems*, *10*, 223-242.

King, W.R., & Teo, T.S.H. (1996). Key dimensions of facilitators and inhibitors for the strategic use of information technology. *Journal of Management Information Systems*, (4), 35-53.

Lertwongsatien, C., & Wongpinunwatana, N. (2003). E-commerce adoption in Thailand: An empirical study of small and medium enterprises (SMEs). *Journal of Global Information Technology Management,* (3), 67-83.

Liu, V., & Khalifa, M. (2003). Determinants of satisfaction at different adoption stages of Internet-based services. *Journal of the Association for Information Systems*, (5), 206-232.

Lockett, N.J., & Brown, D.H. (2000). eClusters: The potential for the emergence of digital enterprise communities enabled by one or more Intermediaries in SMEs. *Journal of Knowledge and Process Management*, (3), 196-206.

McGowan, M.K., & Madey, G.R. (1998). Adoption and implementation of electronic data interchange. In T.J. Larson & E. McGuire (Eds), *Information systems innovation and diffusion: Issues and direction* (pp. 116-140). Hershey, PA: Idea Group Publishing.

Mehrtens, J., Cragg, P., & Mills, A. (2001). A model of Internet adoption by SMEs. *Information and Management*, (3), 165-176.

Pearson, J.M., & Grandon, E. (2004). E-commerce adoption: Perceptions of mangers/owners of small and medium-sized firms in Chile. *Communications of the Association for Information Systems, 13*, 81-102.

Pollard, C. (2003). E-service adoption and use in small farms in Australia: Lessons learned from a government-sponsored programme. *Journal of Global Information Technology Management,* (2), 45-63.

Poon, S., & Swatman, P.M.C. (1997). Small business use of the Internet: Findings from Australian case studies. *International Marketing Review*, (5), 385-402.

Raghunathan, M., & Madey, G.R. (1999). A firm-level framework for planning electronic commerce information systems infrastructure. *International Journal of Electronic Commerce*, (1), 125-145.

Raymond, L. (2001). Determinants of Web site implementation in small businesses. *Internet Research,* (5), 411-422.

Raymond, L., & Bergeron, F. (1996). EDI success in small and medium-sized enterprises: A field study. *Journal of Organizational Computing and Electronic Commerce*, (2), 161-172.

Rogers, E.M. (1991). The critical mass in the diffusion of interactive technologies in organizations. In K.L. Kraemer (Ed.), *The information systems research challenge: Survey research methods* (Vol. 3, pp. 245-217). Boston: Harvard Business School.

Rogers, E.M. (2003). *Diffusion of innovations* (5th edition). New York: The Free Press.

Scupola, A. (2003). The adoption of Internet commerce by SMEs in the south of Italy: An environmental, technological and organizational perspective. *Journal of Global Information Technology Management, 6*(1), 52-71.

Teo, T.S.H., & Tan, M. (1998). An empirical study of adoptors and non-adopters of the Internet in Singapore. *Information and Management*, (6), 339-345.

Teo, T.S.H., & Pian, Y. (2003). A contingency perspective on Internet adoption and competitive advantage. *European Journal of Information Systems*, (2), 78-92.

Teo, T.S.H., & Pian, Y. (2004). A model for Web adoption. *Information and Management,* (4), 457-468.

Thong, J.Y.L. (1999). An integrated model of information systems adoption in small business. *Journal of Management Information Systems*, (4), 187-214.

Walczuch, R., Van Braven, G., & Lundgren, H. (2000). Internet adoption barriers for small firms in The Netherlands. *European Management Journal,* (5), 561-572.

Webb, B., & Sayer, R. (1998). Benchmarking small companies on the Internet. *Long Range Planning*, (6), 815-827.

Wheeler, B.C. (2003). NEBIC: A dynamic capabilities theory for assessing net-enablement. *Information Systems Research*, (2), 125-146.

Zaltman, G., Duncan, R., & Holbek, J. (1973). *Innovations and Organizations*. New York: John Wiley & Sons.

KEY TERMS

Adoption Factors: These are the major factors that encourage (or discourage) an organization to adopt an innovation. These include the following keywords.

Environmental Factors: These reflect pressures to adopt an innovation that are external to the organization. These pressures may be exerted by competitors, clients, trading partners, government initiatives and other characteristics of the marketplace.

Innovation Factors: These reflect characteristics of a specific innovation that encourages an organization to adopt an innovation, including the *perceived benefits* of the innovation and the *compatibility* of the innovation to the organization.

Internet Adoption: Occurs when a firm embraces an Internet application for the first time. Typical Internet applications include e-mail, Web browsing, Web site presence and electronic transactions. Firms will often adopt Internet technology in stages (or levels) over time, beginning with one application and adding another and so on. Each new application can be regarded as an Internet adoption.

Management Support: Managers can provide support for an innovation project. However, this support can be offered in various ways. For example, some managers take the lead role in a project (e.g., as a project champion or project manager), as they are keen to see the organization adopt the innovation. Other managers may adopt a less-direct role; for example, by giving approval for financial expenditure but not getting involved in the project.

Organizational Factors: These include the resources that an organization has available to support the adoption of an innovation, such as financial and technological resources as well as top management support, and top management knowledge.

Small Firm: There is no one universally accepted definition for a small firm. While most definitions are based on the number of employees, some include sales revenue. For example, 20 employees is the official definition in New Zealand, while in North America, a firm with 500 could be defined as a small firm. Another important aspect of any definition of "small firm" is the firm's independence; that is, a small firm is typically considered to be independent and not a subsidiary of another firm.

Internet Privacy from the Individual and Business Perspectives

Tziporah Stern
Baruch College, CUNY, USA

INTRODUCTION: PRIVACY

People have always been concerned about protecting personal information and their right to privacy. It is an age-old concern that is not unique to the Internet. People are concerned with protecting their privacy in various environments, including healthcare, the workplace and e-commerce. However, advances in technology, the Internet, and community networking are bringing this issue to the forefront. With computerized personal data files:

a. retrieval of specific records is more rapid;
b. personal information can be integrated into a number of different data files; and
c. copying, transporting, collecting, storing, and processing large amounts of information are easier.

In addition, new techniques (i.e., data mining) are being created to extract information from large databases and to analyze it from different perspectives to find patterns in data. This process creates new information from data that may have been meaningless, but in its new form may violate a person's right to privacy. Now, with the World Wide Web, the abundance of information available on the Internet, the many directories of information easily accessible, the ease of collecting and storing data, and the ease of conducting a search using a search engine, there are new causes for worry (Strauss & Rogerson, 2002). This article outlines the specific concerns of individuals, businesses, and those resulting from their interaction with each other; it also reviews some proposed solutions to the privacy issue.

CONTROL: PRIVACY FROM THE INDIVIDUAL'S PERSPECTIVE

The privacy issue is of concern to many types of people and individuals from different backgrounds. Gender, age, race, income, geographical location, occupation, and education level all affect people's views about privacy. In addition, culture (Milberg et al., 2000; Smith, 2001) and the amount of Web experience accumulated by an individual is likely to influence the nature of the information considered private (Hoffman et al., 1999; Miyazaki & Fernendez, 2001). Table 1 summarizes the kinds of information people would typically consider private.

When interacting with a Web site, individuals as consumers are now more wary about protecting their data. About three-quarters of consumers who are not generally concerned about privacy fear intrusions on the Internet (FTC, 2000). This is due to the digitalization of personal information, which makes it easier

Table 1. Private information

Information
▪ Address
▪ Credit card numbers
▪ Date of birth
▪ Demographic information
▪ E-mail
▪ Healthcare information and medical records
▪ Name
▪ Phone number
▪ Real-time discussion
▪ Social Security number
▪ Usage tracking/click streams (cookies)

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 2. Individual's concerns

Concerns
▪ Access
▪ Analyzing
▪ Collection
▪ Combining data
▪ Contents of the consumer's data storage device
▪ Creating marketing profiles of consumers
▪ Cross matching
▪ Distributing and sharing
▪ Errors in data
▪ Identity theft
▪ Reduced judgment in decision making
▪ Secondary use of data
▪ Selling data (government)
▪ Spam
▪ Storing
▪ Use
▪ Video surveillance on the Internet
▪ Web bugs

for unauthorized people to access and misuse it (see Table 2 for a list of concerns regarding the uses of data). For example, many databases use Social Security numbers as identifiers. With this information and the use of the Internet, personal records in every state's municipal database can be accessed (Berghel, 2000).

There also are many issues regarding policies and security controls. Individuals are concerned about breaches of security and a lack of internal controls (Hoffman, 2003). However, surprisingly, about one-third of Web sites do not post either a privacy policy or an information practice statement (Culnan, 1999), and only about 10% address all five areas of the Fair Information Practices (FIP), U.S. guidelines to protect computerized information (see FIP in Terms section) (Culnan, 1999; Federal Trade Commission, 2000). Additionally, there is a mismatch between policies and practices (Smith, 2001); this means that a company may publicize fair information policies but in practice does not follow its own guidelines.

Furthermore, as a result of the data mining technology, computer merging and computer matching have become a new privacy concern. One reason is because individuals may have authorized data for one purpose but not for another, and through data mining techniques, this information is extracted for further use and analysis. For example, a consumer's information may have been split up among many different databases. However, with sophisticated computer programs, this information is extracted and used to create a new database that contains a combination of all the aggregate information. Some of these data mining techniques may not be for the benefit of the consumer. It may allow the firms to engage in price and market discrimination by using consumers' private information against them (Danna & Gandy, 2002).

Some additional concerns are whether the Web site is run by a trusted organization, whether individuals can find out what information is stored about them, and whether their name will be removed from a mailing list, if requested. Consumers also want to know who has access to the data and if the data will be sold to or used by third parties. They want to know the kind of information collected and the purpose for which it is collected (Cranor et al., 1999; Hoffman, 2003). In addition, consumers want to feel in control of their personal information (Hoffman, 2003; Olivero & Lunt, 2004). According to a Harris Poll (2003), 69% of consumers feel they have lost control of their personal information.

TRUST: PRIVACY FROM THE BUSINESS PERSPECTIVE

Privacy also is important to businesses. A business collects information about its customers for many reasons: to serve them more successfully, to build a long-term relationship with them, and to personalize services. To build a successful relationship, businesses must address their customers' privacy concerns (Resnick & Montania, 2003) so that their customers will trust them. They must also protect all information they have access to, since this is what consumers expect of them (Hoffman et al., 1999). Furthermore, they must be aware of the fact that some information is more sensitive (Cranor et al., 1999), such as Social Security numbers (Berghel, 2000). This trust is the key to building a valuable relationship with customers (Hoffman et al., 1999; Liu et al. 2004).

One of the many ways a business can gain consumer confidence is by establishing a privacy policy, which may help consumers trust it and lead them to return to the Web site to make more purchases (Liu

et al., 2004). When a business is trusted, consumers' privacy concerns may be suppressed, and they may disclose more information (Xu et al., 2003). Privacy protection thus may be even more important than Web site design and content (Ranganathan & Ganapathy, 2002). Also, if an organization is open and honest with consumers, the latter can make a more informed decision as to whether or not to disclose information (Olivero & Lunt, 2004).

INDIVIDUAL VS. BUSINESS = PRIVACY VS. PERSONALIZATION

In matters of information, there are some areas of conflict between businesses and consumers. First, when a consumer and an organization complete a transaction, each has a different objective. The consumer does not want to disclose any personal information unnecessarily, and a business would like to collect as much information as possible about its customers so that it can personalize services and advertisements, target marketing efforts, and serve them more successfully. Consumers do appreciate these efforts yet are reluctant to share private information (Hoffman, 2003).

Cookies

Second, search engines also may potentially cause privacy problems by storing the search habits of their customers by using cookies. Their caches also may be a major privacy concern, since Web pages with private information posted by mistake, listserv, or Usenet postings may become available worldwide (Aljifri & Navarro, 2004). In general, cookies may be a privacy threat by saving personal information and recording user habits. The convenience of having preferences saved does not outweigh the risks associated with allowing cookies access to your private data. There are now many software packages that aid consumers in choosing privacy preferences and blocking cookies (see solutions section).

Google

Finally, the most recent controversy involves Google's Gmail service and Phonebook. Gmail uses powerful search tools to scan its users' e-mails in order to provide them with personalized advertisements. On the one hand, this invades users' private e-mails. However, it is a voluntary service the user agrees to when signing up (Davies, 2004).

SOLUTIONS

There have been many attempts at trying to solve the privacy problem. There are three different types of solutions: governmental regulation, self-regulation, and technological approaches.

Governmental Regulation

Some form of government policy is essential, since in the absence of regulation and legislation to punish privacy-offenders, consumers may be reluctant to share information. However, written privacy policy requires enforcement (O'Brien & Yasnoff, 1999). In addition, given the current bureaucratic nature of legislation, technology advances far faster than the laws created to regulate it. Consequently, self-regulation may be a better solution.

Self-Regulation

There are numerous forms of self-regulation. The Fair Information Practices (U.S.) and the Organization for Economic Co-operation and Development (OECD) Guidelines (International) are both guidelines for protecting computerized records. These guidelines provide a list of policies a company should follow. Another type of self-regulated solution is a privacy seal program such as TRUSTe or Verisign. A business may earn these seals by following the guidelines that the seal company provides. A third kind of self-regulation is opt-in/opt-out policy. Consumers should be able to choose services by opting in or out (Hoffman et al., 1999) and to voluntarily embrace new privacy principles (Smith, 2001). A joint program of privacy policies and seals may provide protection comparable to government laws (Cranor et al., 1999) and may even address new issues faster than legislation.

Technology

Technological solutions also are a viable alternative. Technologies can protect individuals by using encryption, firewalls, spyware, and anonymous and pseudonymous communication. A well-known privacy technology is the Platform for Privacy Preferences (P3P), a World Wide Web Consortium (W3C) project that provides a framework for online interaction and assists users in making informed privacy decisions. In summary, although there seems to be some promise to each of these three alternatives, a combination of government regulation, privacy policies, and technology may be the best solution.

CONCLUSION

Advances in the collection and analysis of personal information have proven to be beneficial to society. At the same time, they have aggravated the innate concern for the protection of privacy. This article has reviewed current issues in the areas of information privacy and its preservation. It has included the differing points of view of those providing the information and those collecting and using it. Since the collection of information entails both benefits and threats, various suggestions for minimizing the economic costs and maximizing the benefits are discussed.

REFERENCES

Aljifri, H., & Navarro, D.S. (2004). Search engines and privacy. *Computers and Security, 23*(5), 379-388.

Berghel, H. (2000). Identity theft, Social Security numbers, and the Web. *Communications of the ACM, 43*(2), 17-21.

Cranor, L.F., Reagle, J., & Ackerman, M.S. (1999). Beyond concern: Understanding net users' attitudes about online privacy. *AT&T Labs-Research Technical Report TR 99.4.3*. Retrieved April 5, 2004, from *http://www.research.att.com/resources/trs/TRs/99/99.4/99.4.3/report.htm*

Culnan, M.J. (1999). Georgetown Internet privacy policy survey: Report to the Federal Trade Commission. Retrieved April 3, 2004, from *http://www.msb.edu/faculty/culnanm/gippshome.html*

Danna, A., & Gandy Jr., O.H. (2002). All that glitters is not gold: Digging beneath the surface of data mining. *Journal of Business Ethics, 40*(4), 373-386.

Davies, S. (2004). Privacy international complaint: Google Inc.—Gmail email service. Retrieved June 22, 2004, from *http://www.privacyinternational.org/issues/internet/gmail-complaint.pdf*

Federal Trade Commission. (2000). Privacy online: Fair information practices in the electronic marketplace. Retrieved September 23, 2004, from *http://www.ftc.gov/reports/privacy2000/privacy2000.pdf*

Harris Poll. (2003). Most people are "privacy pragmatists" who, while concerned about privacy, will sometimes trade it off for other benefits. Retrieved September 23, 2004, from *http://www.harrisinteractive.com/harris_poll/index.asp?PID=365*

Hoffman, D.L. (2003). The consumer experience: A research agenda going forward. *FTC public workshop1: Technologies for protecting personal information: The consumer experience. Panel: Understanding how consumers interface with technologies designed to protect consumer information*. Retrieved June 6, 2004, from *http://elab.vanderbilt.edu/research/papers/pdf/manuscripts/FTC.privacy.pdf*

Hoffman, D.L., Novak, T.P., & Peralta, M. (1999). Information privacy in the marketplace: Implications for the commercial uses of anonymity on the Web. *The Information Society, 15*(2), 129-140.

Liu, C., Marchewka, J.T., Lu, J., & Yu, C.S. (2004). Beyond concern—A privacy-trust—Behavioral intention model of electronic e-commerce. *Information & Management, 42*(1), 127-142.

Milberg, S.J., Smith, H.J., & Burke, S.J. (2000). Information privacy: Corporate management and national regulation. *Organization Science, 11*(1), 35-58.

Miyazaki, A.D., & Fernandez, A. (2001). Consumer perceptions of privacy and security risks for online

shopping. *The Journal of Consumer Affairs, 35*(1), 27-55.

O'Brein, D.G., & Yasnoff, W.A. (1999). Privacy, confidentiality, and security in information systems of state health agencies. *American Journal of Preventive Medicine, 16*(4), 351-358.

Olivero, N., & Lunt, P. (2004). Privacy versus willingness to disclose in e-commerce exchanges: The effect of risk awareness on the relative role of trust and control. *Journal of Economic Psychology, 25*(2), 243-262.

Ranganathan, C., & Ganapathy, S. (2002). Key dimensions of business-to-consumer Web sites. *Information & Mangement, 39*(6), 457-465.

Smith, H.J. (2001). Information privacy and marketing: What the US should (and shouldn't) learn from Europe. *California Management Review, 43*(2), 8-34.

Strauss, J., & Rogerson, K.S. (2002). Policies for online privacy in the United States and the European Union. *Telematics and Informatics, 19*(2), 173-192.

Xu, Y., Tan, B.C.Y., Hui, K.L., & Tang, W.K. (2003). Consumer trust and online information privacy. *Proceedings of the Twenty-Fourth International Conference on Information Systems*, Seattle, Washington.

KEY TERMS

Cookies: A string of text that a Web browser sends to you while you are visiting a Web page. It is saved on your hard drive, and it saves information about you or your computer. The next time you visit this Web site, the information saved in this cookie is sent back to the Web browser to identify you.

Data Mining: A process by which information is extracted from a database or multiple databases using computer programs to match and merge data and create more information.

Fair Information Practices (FIP): Developed in 1973 by the U.S. Department of Health, Education, and Welfare (HEW) to provide guidelines to protect computerized records. These principles are collection, disclosure, accuracy, security, and secondary use. Some scholars categorize the categories as follows: notice, choice, access, security, and contact information (Culnan, 1999; FTC, 2000).

Opt-In/Opt-Out: A strategy that a business may use to set up a default choice (opt-in) in a form that forces a customer, for example, to accept e-mails or give permission to use personal information, unless the customer deliberately decline this option (opt-out).

Organization for Economic Co-operation and Development (OECD) Guidelines: International guidelines for protecting an individual's privacy (similar to the FIP).

Privacy: The right to be left alone and the right to control and manage information about oneself.

Privacy Seals: A seal that a business may put on its Web site (i.e., Verisign or TRUSTe) to show that it is a trustworthy organization that adheres to its privacy policies.

Internet Privacy Issues

Hy Sockel
Youngstown State University, USA

Kuanchin Chen
Western Michigan University, USA

WHAT IS INTERNET PRIVACY?

Businesses need to understand privacy conditions and implications to ensure they are in compliance with legal constraints and do not step on consumers' rights or trust. Personal identifiable information (PII) and data can have innate importance to an organization. Some organizations view certain privacy features as essential components of their products or services; for example, profile data is often used to tailor products specifically for their customers' likes and needs. PII can also be used for less honorable endeavors, such as identity theft, phishing, political sabotage, character annihilation, spamming and stalking.

One of the core issues of privacy is who actually owns the data—the holder of it, or the persons that the data is about? The answer depends on many criteria: the users' perspective, the environment in which that privacy is addressed, and how the data are collected and used. Privacy issues arise because every Internet transaction leaves an important artifact of every transaction the individual did when searching for information, shopping or banking. This audit trail has caused many people to be concerned that this data may be inappropriately used. The paradox is that many businesses are also concerned. They believe that government, in its haste to protect individuals' privacy, could interfere with the development of new services, technologies and the electronic marketplace.

It is important to state that the government's approach to the protection of personal privacy is neither equal nor universal. Some localities extend protection much further than others. In 1972, California amended its constitution to specifically include the construct of "a resident's inalienable right to privacy." Within the United States (U.S.), court decisions dealing with privacy have fairly closely upheld two principles (Freedman 1987):

1. The right to privacy is NOT an absolute. An individual's privacy has to be tempered with the needs of society.
2. The public's right to know is superior to the individual's right of privacy.

VIOLATION OF PRIVACY AS AN UNACCEPTABLE BEHAVIOR

The Internet Activities Board (IAB) issued a Request for Comment (RFC-1087) in 1989 dealing with what they characterized as the proper use of Internet resources. Prominent on the IAB's list of what it considers as unethical and unacceptable Internet behavior is the act that "compromises the privacy of users." The reliable operation of the Internet and the responsible use of its resources are of common interest and concern for its users, operators and sponsors (Stevens, 2002).

Using the Internet to violate people's privacy by targeting them for abusive, corrosive comments or threats is not only unacceptable, but illegal. Privacy violations can do a lot more than just embarrass individuals. Information can be used in blackmail or otherwise coerce behavior. Institutions could use information to deny loans, insurance or jobs because of medical reasons, sexual orientation or religion. People could lose their jobs if their bosses were to discover private details of their personal life.

ONLINE PRIVACY AND DATA COLLECTION

Online privacy concerns arise when PII is collected online without the consumers' consent or knowledge and is then disseminated without the individual's "blessing." Dhillon and Moores (2001) found that the

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

top-five list of Internet privacy concerns include (a) personal information sold to others; (b) theft of personal data by a third party; (c) loss of personal files; (d) hacker's damage to personal data; and (e) spam. Cockcroft (2002) suggested the following top privacy concerns: (a) unauthorized secondary use; (b) civil liberties; (c) identity theft; (d) data profiling; and (e) unauthorized plugins. Online privacy is generally considered as the right to be left alone and the right to be free from unreasonable intrusions. By extrapolation, one can label telemarketers, mass advertisements, "spam," online "banner ads" and even commercials to be relating directly to privacy issues because of the solitude and intimacy dimension. Westin (1970) frames privacy into four dimensions:

a. **Solitude:** the state of being alone away from outside interference.
b. **Intimacy:** the state of privacy one wants to enjoy from the outside world.
c. **Anonymity:** the state of being free of external surveillance.
d. **Reserve:** the ability to control information about oneself.

While organizations can go the "extra mile" to safeguard data through the data collection, transmission and storage processes, this may not be sufficient to keep client content private. Some businesses use the collected user information for credit worthiness checks, mass customization, profiling, convenience, user tracking, logistics, location marketing and individualized services. The issue sometimes breaks down to whom has more rights to control the data:

a. the organization that committed resources to collect and aggregate the data; or
b. the people the data is about.

When information is collected, there is the matter of trust: Consumers have to decide if they trust the organization to use the data appropriately. The organization has to trust that the information they asked for represents the facts.

Violating privacy hurts everyone. If people no longer believe their data will be handled appropriately, there is less incentive for them to be honest. "Almost 95% of Web users have declined to provide personal information to Web sites at one time or another when asked" (Hoffman et al., 1999, p. 82). Of those individuals that do provide information, more than half of them have admitted to lying on collection forms and in interviews. Chen and Rea (2004) indicated that concern of unauthorized information use is highly related to passive reaction. Passive reaction is one type of privacy control, where one simply ignores data collection requests. Users tend to exercise another privacy control—identity modification—when they are highly concerned about giving out personal information for any reason.

ACTIVITIES THAT MAY VIOLATE PERSONAL PRIVACY

Cookies and Web-Bugs

A cookie is a small amount of information that the Web server requests the user's browser to save on the user's machine. Cookies provide a method of creating persistent memory for an organization in the stateless environment of the native Internet. Organizations use cookies to collect information about the users and their online activities to "better serve" their clients, but some go beyond the honest use of cookies by involving third parties to also plant their cookies on the same Web page. The collected information about the users may be resold or linked to external databases to form a comprehensive profile of the users. Web-bugs (or clear images) allow for user tracking, but they can easily go unnoticed. Most browsers give the user an option to deny or allow cookies, but very few of them are capable of filtering out Web-bugs.

Spam

Any time users enter their e-mail address on a Web site, they run the risk of being added to an e-mail list. The e-mail address is often packaged and sold to merchants. In the end, users end up being bombarded with unwanted and often offensive e-mails. Spam is a pervasive problem in the wired world; automated technologies can send e-mails by the hundreds of thousands. Spam taxes Internet servers, annoys con-

sumers and is an abuse of an intended system. To thwart spam, privacy advocates with only moderate success battle with Internet Service Providers (ISPs), e-mail providers and Internet application providers.

Spoofing and Phishing

One major concern on the Internet is ensuring that users are dealing with who they think they are. Spoofing is the act to deceive; in the Internet world, it is the act of pretending to be someone by fooling the hardware, software or the users. Even when a user lands on what appears to be a familiar site, not everything is as it appears to be. Thieves have usurped legitimate Web sites' look and feel in a process known as "Phishing." The phishing scam requests users to supply personal identification information so they may be verified. The Web thieves take the supplied information—unbeknownst to the user—and then respond in a fashion that makes everything appear normal. Some Web sites have employed digital certificates to try to battle hackers and phishing schemes. Although browsers automatically verify the legitimacy of certificates, they cannot tell the users that the Web site (with a legitimate certificate) is indeed what the users intend to visit.

PRIVACY: CURRENT PRACTICE

Information Opt-In and Opt-Out

A major debate exists over how an organization should acquire user consent. At the heart of the privacy debate is the tug-of-war between those in favor of opt-in vs. opt-out policies. The "Opt-In" group believes that organizations should be forced to individually seek consent from each user each time they collect data from the user. The "Opt-Out" group finds it perfectly acceptable, by default, to include everyone (and their data) and force the users to deny consent. The Opt-Out is the most common form of policies in the U.S. The effectiveness of the Opt-Out notice is questionable, because the notice is typically written in a legalese fashion. For the average user, it is typically vague, incoherent and intentionally hidden in verbose agreements.

Another privacy issue revolves around data about an individual that comes into the possession of a "third party." According to the Opt-In group, the individual should be able to exercise a substantial degree of control over that data and its use (Clarke, 1998). The issue of control takes on major significance with the U.S. and the European community on different sides of the Opt-In and Opt-Out discussion. The U.S. government has codified the Opt-Out requirement under several different acts, such as Gramm-Leach-Bliley (GLB) Act and The Fair and Accurate Credit Transactions Act of 2003 (FACT).

Privacy Impact Assessment (PIA)

PIAs are becoming popular instruments to ensure compliance with appropriate industrial, organizational and legal guidelines. PIAs became mandatory in Canada as of 2002. Basically, PIAs are proactive tools that look at both policy and technology risks to ascertain the effects of initiatives on individual privacy.

In practice, PIAs tend to be primarily "policy focused" and rarely address the underlying information management and technology design issues. Consequently, PIAs tend to "blur and not cure" the issue of personal information misuse. Worse yet, PIAs can mislead organizations into false senses of security. Organizations may feel they are compliant with applicable regulations because they post a privacy policy on their Web site. However, many privacy notifications, even if they are 100% guaranteed to be delivered, do not address the issues of compliance in data collection, data handling and secondary use of PII.

Privacy Seals

The lack of effectiveness on the part of the government to adequately address protection of consumer privacy has caused the rise of privacy advocate organizations. These "privacy organizations" (such as BBBOnline and TRUSEe) inspect Web site privacy policies and grant a "seal of approval" to those who comply with industry privacy practice. However, because these organizations "earn their money from e-commerce organizations, they become more of a privacy advocate for the industry—rather than for the consumers" (Catlett, 1999).

Some groups argue that seals of any kind are counter-productive. A consumer visiting a site may develop a false sense of security, which could be worse than knowing the data submitted to the site is insecure. Even if the seal on an organization's Web site was legitimately acquired, there is no guarantee that the organization still follows the same procedures and policies they used after they acquired the seal. An organization may change its attitudes over time, may not keep its privacy statements up to date, or may even change its privacy statements too frequently. Unfortunately, there is no real mechanism to know that a site changed its policies after it acquired a seal. A very troublesome assumption of privacy seals presupposes that users of a Web site review the privacy statement and understand its legal implications each time they use that Web site.

Platform for Privacy Preferences (P3P) Project

The industry has taken a number of steps, through privacy seal programs and self-regulatory consortiums, to adopt standards to protect online privacy. The World Wide Web Consortium (W3C) has contributed to this effort with the P3P. According to the W3C Web site (*www.w3.org/P3P*), P3P "is a standardized set of multiple-choice questions, covering all the major aspects of a Web site's privacy policies." P3P-enabled Web sites work with P3P-enabled browsers to automatically handle the users' personal information according to the set of personal privacy preferences.

The idea is quite eloquent in its simplicity, which is to put privacy policies where users can find them, in ways users can understand and that users can control. While the W3C purports P3P to be a simple approach to privacy protection, it fails to address one of the core problems of privacy statements: the legalese. The legal language that many Web sites' privacy policy statements are written in bewilders many users (Zoellick, 2001).

Digital Certificates

An entirely different approach to privacy and authentication uses certificates instead of "seals of approval." In the physical realm, a certificate might be someone's signature or a communication of some kind (document, letter or verbal) from a known friend or trusted colleague to attest to another person's identity, skills, value or character. As such, the person giving the "communicate" lends credence to another party (person, group or organization).

The electronic equivalent of an individual's signature or that of a trusted "communiqué" involves digital certificates—a unique digital ID used to identify individuals. Digital certificates are based on a hierarchy of trust. At the top level of the hierarchy needs to be a well-trusted root entity. Off of the root entity trust is disseminated downwards, with each new level being verified by the level above it.

Cryptography and the Law: "Wassenaar Arrangement"

A complex web of national and international laws regulate the use of cryptography. Thirty-three countries joined together to form the "Wassenaar Arrangement" group. The group's goal is to make uniform decisions on the export of "dual use" technology, such as cryptography. Participating members seek, through their national policies, "to ensure that transfers of dual-use items do not contribute to the development or enhancement of military capabilities, which undermine these goals, and are not diverted to support such capabilities" (*www.wassenaar.org*). According to the Arrangement, the decisions on transfer or denial of transfer of any item are the sole responsibility of each member country (Madsen & Banisar, 2000). In the U.S., it is illegal to export strong cryptographic software. In other countries, such as France, any use of strong encryption is forbidden.

European Union Privacy Laws

The European Union (EU) supports very strong consumer privacy standards. The EU's comprehensive privacy legislation, the "Directive on Data Protection," became effective October 25, 1998. The Directive requires that transfers of personal data take place only to non-EU countries that provide an "adequate" level of privacy protection. The problem is that a large amount of U.S. companies facing the Directive stringent mandates use a mix of legislation, regulation and self-regulation, which do not satisfy all the EU's requirements. Specifically, under the

Directive's consumers must have access to the information stored about them so they can correct erroneous data. Because many U.S. companies cannot fulfill this requirement, the exchange of data across international borders is problematic. A "Safe Harbor" arrangement has been reached.

Millennium ACT - EU

The European Union Copyright Directive (EUCD) and the U.S. Digital Millennium Copyright Act (DMCA) are both, in part, modeled after the World Intellectual Property Organization (WIPO) Copyright Treaty and the WIPO Performances and Phonogram Treaty. Sony Corp. filed a lawsuit under the Italian version of the EU equivalent of the DMCA (passed April 2003) addressing people purchasing the modified PlayStation. The lawsuit had local authorities confiscate the modified game systems as a violation of the EUCD. On December 31, 2003, the Italian court declared the seizures illegal. The court ruled that the new law did not apply because the chips in question were not intended primarily to circumvent copyright protection measures.

U.S.A. Patriot Act of October 11, 2001

On October 11, 2001, President Bush signed into law the *Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism* Act, better known as the U.S.A. Patriot Act. Under the guise of the Patriot Act, two very controversial programs were authorized: DCS1000 (a.k.a. Carnivore), and Total Information Awareness (TIA). The Patriot Act permits the FBI to use technology for monitoring e-mail and other communication. The TIA is a very controversial project directed by the Information Awareness Office (IAO). The IAO's goal is to gather intelligence on possible terrorist activities through electronic sources, such as the Internet, and telephone and fax lines. Many privacy advocates are very concerned that privacy will take a back seat to patriotism and efforts to stamp out terrorism. Certainly, it is essential to provide law enforcement with the means necessary to track down terrorist activity in any medium, but this must be done within a system of expeditious checks and balances.

CAN-SPAM

On December 16, 2003, President Bush signed into law the Controlling the Assault of Non-Solicited Pornography and Marketing Act (CAN-SPAM). The Act establishes a framework to help America's consumers, businesses and families combat unsolicited commercial e-mail, known as spam. The CAN-SPAM is an opt-out law. While recipient permissions are not required to send an e-mail, failure of the organization to abide by a recipient's desire to opt-out carries penalties of a fine and/or imprisonment for up to five years and may cause the perpetrators to lose any assets purchased with funds from such an endeavor.

While some claim that this Act finally gives the law enforcement community teeth, others profoundly disagree. SPAMHAUS (2003) indicates the Act is backed overwhelmingly by spammers and has been dubbed the "YOU-CAN-SPAM" Act. They claim that it legalizes spam instead of banning it. The Act, unfortunately, pre-empts state laws that are stronger to protect consumers from being spammed.

CONCLUSION

Although privacy has been addressed in various articles for well more than three decades, new privacy issues continue to emerge along with the introduction of new technology. Readers interested in specific areas of Internet privacy may want to read the cited references, such as Smith (2003) for related legislation; Westin (1970) for early views; and Smith, Milberg and Burke (1996) and Stewart and Segars (2002) for empirical assessments.

REFERENCES

Catlett, J. (1999). 1999 comments to the Department of Commerce and Federal Trade Commission. Retrieved June 10, 2004, from *www.junkbusters.com/profiling.html*

Chen, K., & Rea, A., Jr. (2004). Protecting personal information online: A survey of user privacy concerns and control techniques. *Journal of Computer Information Systems*, forthcoming.

Clarke, R. (1998). Direct marketing and privacy (version of February 23, 1998). Retrieved June 10, 2004, from *www.anu.edu.au/people/Roger.Clarke/DV/DirectMkting.html*

Cockcroft, S. (2002). Gaps between policy and practice in the protection of data privacy. *Journal of Information Technology Theory and Application, 4*(3), 1-13.

Dhillon, G.S., & Moores, T.T. (2001). Internet privacy: interpreting key issues. *Information Resources Management Journal, 14*(4), 33-37.

Freedman, W. (1987). *The right of privacy in the computer age.* New York: Quorum Books.

Hoffman, D.L., Novak, T.P., & Peralta, M. (1999). Building consumer trust online. *Communications of the ACM, 42*(4), 80-85.

Madsen, W., & Banisar, D. (2000). Cryptography and liberty 2000 – An international survey of encryption policy. Retrieved June 10, 2004, from *www2.epic.org/reports/crypto2000/overview*

Smith, H.J., Milberg, S.J., & Burke, S.J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly*, 167-196.

Smith, M.S. (2003). Internet privacy: Overview and pending legislation. Retrieved June 10, 2004, from *www.thememoryhole.org/crs/RL31408.pdf*

SPAMHAUS. (2003). United States set to legalize spamming on January 1, 2004. Retrieved June 10, 2004, from *www.spamhaus.org/news.lasso?article=150*

Stevens, G.M. (2002). CRS Report for Congress online privacy protection: Issues and developments. Retrieved June 10, 2004, from *www.thememoryhole.org/crs/RL30322.pdf*

Stewart, K.A., & Segars, A.H. (2002). An empirical examination of the concern for information privacy instrument. *Information Systems Research, 13*(1), 36-49.

Westin, A. (1970). *Privacy and freedom.* New York: Atheneum.

Zoellick, B. (2001). *CyberRegs: A business guide to Web property, privacy, and patents.* Addison-Wesley.

KEY TERMS

Cookie: A small amount of information that the Web site server requests the user's browser to save on the user's machine.

Digital Certificate: A unique digital ID used to identify individuals (personal certificates), software (software certificates) or Web servers (server certificates). They are based on a hierarchy of trust.

Phishing: A form of spoofing, where users are tricked into providing personal identification information because thieves have stolen the "look and feel" of a legitimate site.

Privacy Impact Assessments (PIA): Proactive tools that look at both the policy and technology risks and attempt to ascertain the effects of initiatives on individual privacy.

Privacy Seals: A third party "icon" that indicates they have inspected the Web site privacy policies and found them NOT to be out of line with the industry.

Spam: Unsolicited communications, typically e-mail, that are unwanted and often offensive.

Spoofing: The act to deceive. In the Internet world, it is the act of pretending to be someone or something else by fooling hardware, software or human users.

Interoperable Learning Objects Management

Tanko Ishaya
The University of Hull, UK

INTRODUCTION

The sharing and reuse of digital information has been an important computing concern since the early 1960s. With the advent of the World Wide Web (from now on referred to as the Web), these concerns have become even more central to the effective use of distributed information resources. From its initial roots as an information-sharing tool, the Web has seen exponential growth in a myriad of applications, ranging from very serious e-business to pure leisure environments. Likewise, research into technology support for education has quickly recognised the potential and possibilities for using the Web as a learning tool (Ishaya, Jenkins, & Goussios, 2002). Thus, Web technology is now an established medium for promoting student learning, and today there are a great many online learning materials, tutorials, and courses supported by different learning tools with varying levels of complexity. It can be observed that there are many colleges and universities, each of which teaches certain concepts based on defined principles that remain constant from institution to institution. This results in thousands of similar descriptions of the same concept. This means that institutions spend a lot of resources producing multiple versions of the same learning objects that could be shared at a much lower cost. The Internet is a ubiquitous supporting environment for the sharing of learning materials. As a consequence, many institutions take advantage of the Internet to provide online courses (Ishaya et al.; Jack, Bonk, & Jacobs, 2002; Manouselis, Panagiotu, Psichidou, & Sampson, 2002). Many other agencies have started offering smaller and more portable learning materials defined as learning objects (Harris, 1999; PROMETEUS, 2002). While there are many initiatives for standardising learning technologies (Anido, Fernandez, Caeiro, Santos, Rodriguez, & Llamas, 2002) that will enable reuse and interoperability, there is still a need for the effective management, extraction, and assembling of relevant learning objects for end-user satisfaction.

What is required, therefore, is a mechanism and infrastructure for supporting a centralized system of individual components that can be assembled according to learners' requirements.

The purpose of this paper is to examine current approaches used in managing learning objects and to suggest the use of ontologies within the domain of e-learning for effective management of interoperable learning objects. In the next section, a background of this paper is presented. The current state of e-learning metadata standards is examined and a brief overview of the semantic-Web evolution in relation to e-learning technology development is given. Then, the paper discusses the driving force behind the need for effective management of interoperability of learning objects. Next, the paper presents e-learning ontologies as the state-of-the-art way of managing interoperable learning objects. Finally, the paper concludes with further research.

BACKGROUND

The background of this paper is based on two different disciplines: developments in Web-based educational systems and the evolving vision of the semantic Web by Berners-Lee et al. (2001).

Web-Based Educational Systems

Electronic learning has been defined as a special kind of technology-based learning (Anderson, 2000; Gerhard & Mayr, 2002). E-learning systems and tools bring geographically dispersed teams together for learning across great distances. It is now one of the fastest growing trends in computing and higher education. Gerhard and Mayr identified three major trends as internalization, commercialization and modularization, and virtualization. These trends are

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

driven by the convenience, flexibility, and time-saving benefits e-learning offers to learners. It is a cost-effective method of increasing learning opportunities on a global scale. Advocates of e-learning claim innumerable advantages ranging from technological issues and didactics to the convenience for students and faculty (Gerhard & Mayr, 2002; Hamid, 2002). These result in tremendous time and cost savings, greatly decreased travel requirements, and faster and better learning experiences. These systems are made possible by the field of collaborative computing (Ishaya et al., 2002), encompassing the use of computers to support the coordination and cooperation of two or more people who attempt to perform a task or solve a problem together. All these seem a promise toward changing how people will be educated and how they might acquire knowledge.

In order to support the increasing demand for Web-based educational applications, a number of virtual learning environments (VLEs) and managed learning environments (MLEs) have since been launched on the market. These VLEs (e.g., Blackboard and WebCT) are a new generation of authoring tools that combines content-management facilities with a number of computer-mediated communication (CMC) facilities, as well as teaching and learning tools. VLEs are learning-management software systems that synthesize the functionality of computer-mediated communications software (e-mail, bulletin boards, newsgroups, etc.) and online methods of delivering course materials. They "have been in use in the higher education sector for several years" and are growing in popularity (MacColl, 2001, p. 227). VLEs began on client software platforms, but the majority of new products are being developed with Web platforms (MacColl). This is due to the expense of client software and the ease of providing personal computers with Web browsers. Furthermore, using the Web as a platform allows the easier integration of links to external, Web-based resources.

Alongside evolutionary representation formats for interoperability, many metadata standards have also merged for describing e-learning resources. Amongst others are learning-object metadata (LOM), the shareable content object reference model (SCORM), the Alliance of Remote Instructional Authoring and Distribution Networks for Europe (ARIADNE), and the Instructional Management System (IMS).

All these metadata models define how learning materials can be described in an interoperable way. The IEEE LOM standard, developed by the IEEE Learning Technology Standards Committee (LTSC) in 1997, is the first multipart standard for learning object metadata consisting of the following.

- **IEEE 1484.12.1:** IEEE Standard for Learning Object Metadata. This standard specifies the syntax and semantics of learning-object metadata, defined as the attributes required to fully and adequately describe a learning object.
- **IEEE 1484.12.2:** Standard for ISO/IEC 11404 Binding for Learning Object Metadata Data Model.
- **IEEE 1484.12.3:** Standard for XML Binding for Learning Object Metadata Data Model.
- **IEEE 1484.12.4:** Standard for Resource Description Framework (RDF) Binding for Learning Object Metadata Data Model.

This standard specifies a conceptual data schema that defines the structure of metadata instances for a learning object.

The LOM standards focus on the minimal set of attributes needed to allow these learning objects to be managed, located, and evaluated. Relevant attributes of learning objects to be described include the type of object, author, owner, terms of distribution, and format (http://ltsc.ieee.org/wg12/). Where applicable, LOM may also include pedagogical attributes such as teaching or interaction style, grade level, mastery level, and prerequisites. It is possible for any given learning object to have more than one set of learning-object metadata. LTSC expects these standards to conform to, integrate with, or reference existing open standards and work in related areas. While, most of these approaches provide a means for describing, sharing, and reusing resources, the concept of interoperability and heterogeneous access to content chunks is yet to be fully achieved.

The Semantic Web

E-learning systems are made possible by the ubiquity of Internet standards such as TCP/IP (transmission-control protocol/Internet protocol), HTTP (hypertext transfer protocol), HTML (hypertext markup language, and XML (extensible markup language), an evolved representation format for interoperability. Additionally, emerging schema and semantic standards, such as XML schema, RDF and its extensions, and the DARPA (Defense Advanced Research Projects Agency) agent markup language and ontology inference layer (DAML + OIL), together provide tools for describing Web resources in terms of machine-readable metadata. This aims at enabling automated agents to reason about Web content and produce intelligent responses to unforeseen situations.

Two of these technologies for developing the semantic Web are already mature and in wide use. XML (http:www.w3.org/XML) lets everyone create their own tags that annotate Web pages or sections of text on a page. Programs can make use of these tags in sophisticated ways, but the programmer has to know what the page writer uses each tag for. So, XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean (Erdmann & Studer, 2000). The meaning of XML documents is intuitively clear due to markups and tags, which are domain terms. However, computers do not have intuition. Tag names per se do not provide semantics. Both data-type definitions (DTDs) and XML schema are used to structure the content of documents but not the appropriate formalism to describe the semantics of an XML document. Thus, XML lacks a semantic model; it has only a tree model, but can play an important role in transportation mechanisms.

The resource description framework (http://www.w3.org/RDFs) provides means for adding semantics to a document. It is an infrastructure that enables the encoding, exchange, and reuse of information-structured metadata. The RDF + RDF schema offers modeling primitives that can be extended according to the needs. RDF also suffers from the lack of formal semantics for its modeling primitives, making interpretation of how to use them properly an error-prone process. Both XML and RDF have been touted as standard Web ontology languages, but they both suffer from expressive inadequacy (see Horrocks, 2002), that is, the lack of basic modeling primitives and the use of poorly defined semantics.

A third technology is the ontology representation languages. Several ontology representation languages and tools are now available—some in their early stages of development—in particular, the Web ontology language (OWL), the W3C (World Wide Web Consortium) recommendation for ontology language. However, DAML, OIL, and DAML + OIL are being used (Fensel, Horrocks, van Harmelen, McGuinness, & Patel-Schneider, 2001). All of these rely on RDF, the subject-predicate-object model, which provides a basic but extensible and portable representation mechanism for the semantic Web. Although ontology representation languages for the semantic Web are in early stages of development, it is fair to say that ontology specification would play an important role in the development of interoperable learning objects. This way, both producer and consumer agents can reach a shared understanding by exchanging ontologies that provide an agreed vocabulary.

DRIVING FORCES

Despite intensive developments in the area of Web-based learning technology and the wide variety of software tools available from many different vendors (e.g., WebCT, Blackboard, AudioGraph), there is increasing evidence of dissatisfaction felt by both instructors and learners (Jesshope, 1999; Jesshope, Heinrich, & Kinshuk, 2000). One of the causes of this dissatisfaction is that these software applications are not able to share learning resources with each other. There is evidence that the future growth of Web-based learning may well be constrained on three fronts: first, dissatisfaction with Web learning resources from students due to a lack of pedagogical underpinning in the design of existing Web learning materials (Govindasmy, 2002); second, the lack of standardisation of learning metadata schemas and course structures (Koper, 2002); and third, the lack of software interfaces that provide interoperability.

Lack of Pedagogical Consideration in the Design of Web-Based Learning Systems

Although the Internet has proved its potential for creating online learning environments to support education (Appelt, 1997; Berners-Lee, 1999; Fetterman, 1998; Harris, 1999; Jack et al., 2002), the full potential of the Internet for transforming education is only just being tapped.

The need to link pedagogy to the prevailing technological infrastructure for Web-based learning was highlighted by Ishaya et al. (2002), Koper (2001), and Mergendoller (1996). They emphasized the need for additional frameworks for Web-based learning. In answer to this requirement, several researchers have offered frameworks for learner-centred Web instruction (Bonk, Kirkley, Hara, & Dennen, 2001; Jack et al., 2002), the integration of the Web in one's instruction, the role of the online instructor (Bonk et al.), and the types and forms of interaction made possible by the emergence of the Web (Jack et al.). The need and potential use of Web agents (Jennings, 2000; Wooldridge, 1997) to support students' learning process by enabling an interactive Web-based learning paradigm has also been identified in Ishaya et al. and Jack et al. There is still evidence that pedagogical issues are neglected within the design of most e-learning systems. This may result in these systems failing due to teachers' reluctance to incorporate their learning resources into those systems, learners avoiding e-learning situations, and the poor performance of learners who do use the systems (Deek, Ho, & Ramadhan, 2001; Govindasmy, 2002; Hamid, 2002; Koper, 2001, 2002). There is also evidence of the lack of consideration for users with learning difficulties in current Web-based learning environments (Koper, 2001; Manouselis et al., 2002). Most of the existing Web-based learning frameworks and models are at the theoretical level and address specific aspects of learning pedagogy (e.g., Bonk et al.; Ishaya et al.; Jack et al.).

Lack of Interoperability and Shareable Learning Objects

A wide variety of teaching materials have been made available in a number of specific formats that are no longer supported (Deek et al., 2001; Koper, 2002). These materials are therefore no longer usable without large investments in converting them into a usable format. The reusability of educational content and instructional components is often limited because existing components cannot easily be obtained for integration. The reusability of learning components involves a number of processes such as the identification of components, correct handling of intellectual property rights, isolation, decontextualisation, and the assembly of components (Koper, 2001, 2002). Making components reusable and manageable provides the advantage of efficiency in Web-based learning-system design. The technique, however, is not simple and requires clear agreements about the standards to be used. Software reuse is a key aspect of good software engineering. One of the current trends in this field is the component-based approach (Lim, 1998). Enterprise JavaBeans (EJB) and the common request broker architecture (CORBA) are examples of technologies that are based on the software-component concept. Software reuse allows programmers to focus their efforts on the specific business logic. The component-based software-engineering approach can be used to provide interoperable and shareable learning objects.

Learning-technology standardisation is taking the lead role in the research efforts surrounding Web-based education. Standardisation is needed for two main reasons. First, educational learning resources are defined, structured, and presented using different formats. Second, the functional modules that are embedded in a particular learning system cannot be reused by another one. Projects like IEEE's LTSC (IEEE, 2002), IMS (IEEE), PROMETEUS (2002), GESTALK (1998), and many others are contributing to this standardisation process. The IEEE LTSC is the institution that is gathering recommendations and proposals from other learning-standardisation institutions and projects.

Lack of Industry Guidance for the Design of Manageable Systems

Industry and academic reports highlight the importance of defining metadata for learning (Anido et al., 2002; IEEE, 2002; Koper, 2002). Its purpose is to facilitate and automate the search, evaluation, ac-

quisition, and use of Web-based learning resources. The result so far is the LOM specification (IEEE) proposed by IEEE LTSC, which is becoming a de facto standard.

Personalisation is increasingly being used in e-commerce as an aid to customer-relationship management (CRM) to provide better service by anticipating customer needs. This is because companies believe that this will make interaction more satisfying. In the educational sector, the aim is toward ensuring that Web resources improve students' learning process. This, too, could be improved through personalisation. The semantic Web offers the possibility of providing the user with relevant and customised information (Berners-Lee, 1999). Furthermore, the recognition of the key role that ontologies are likely to play in the future of the Web has lead to the extension of Web markup languages in order to facilitate content description and the development of Web-based ontologies, for example, the XML schema (Horrocks & Tessaris, 2002), RDF (Horrocks & Tessaris; IEEE, 2002), and the recent DAML + OIL (IEEE). While the development of the semantic Web and of Web ontology languages still presents many challenges, it provides a means for creating a centralized and managed Web-based learning environment where software agents (Wooldridge, 1997) can be designed to carry out sophisticated tasks for users. This will provide an adaptive learning environment.

This brief review highlights the complexity of the factors influencing the effectiveness of Web-based learning. Despite the extent of the work mentioned above, there is a lack of an effective way of managing centralized and interoperable learning materials. Some work has addressed the content and sequencing of learning objects (Koper, 2002). However, without a comprehensive pedagogical analysis in the area of Web-based learning, it is difficult to develop learning resources that can be interoperable, interactive, and collaborative. The progress made in understanding and building flexible and interoperable subject-domain and course ontologies, and linking them with learning materials and outcomes has being the emphasis in recent research. Recent developments related to the semantic Web (Berners-Lee, 1999; Horrocks, 2002; Horrocks & Tessaris, 2002) and ontologies (Horrocks) have revealed new horizons for defining structures for authoring interoperable learning objects. This indicates that the models and frameworks drawn will have to be evaluated across different scenarios of use, which should be based on sound software engineering and learning pedagogy.

ONTOLOGIES: A WAY FORWARD

Ontology is not a new concept. The term has a long history of use in philosophy, in which it refers to the subject of existence and particularly a systematic account of existence (Erdmann & Studer, 2000; Gruber, 1995). It has been a co-opted term from philosophy used in computing to describe formal, shared conceptualizations of a particular domain (Gruber). Ontologies have become a topic of interest in computer science (Fensel et al., 2001). An ontology represents information entities such as people, artifacts, and events in an abstract way. They allow the explicit specification of a domain of discourse, which permits access to and reason about agent knowledge (Erdmann & Studer). Ontologies are designed so that knowledge can be shared with and among people and possibly intelligent agents. Tom Gruber defines ontology as "an explicit representation of a conceptualisation. The term is borrowed from philosophy, where Ontology is a systematic account of existence. For AI [artificial intelligence] systems, what 'exists' is that which can be represented" (p. 911).

A conceptualization refers to an abstract model of some phenomenon in the world made by identifying the relevant concept of that phenomenon. Explicit means that the types of concepts used and the constraints on their use are explicitly defined. This definition is often extended by three additional conditions. The fact that an ontology is an explicit, formal specification of a shared conceptualization of a domain of interest indicates that an onotology should be machine readable (which excludes natural language). It indicates that it captures consensual knowledge that is not private to an individual, but accepted as a group or committee of practice. The reference to a domain of interest indicates that domain ontologies do not model the whole world, but rather model just parts that are relevant to the task at hand.

Ontologies are therefore advanced knowledge representations that consist of several components including concepts, relations and attributes, instances, and axioms. Concepts are abstract terms that are organized in taxonomies. Hierarchical concepts are linked with an "is a" relation. For example, we can define two concepts: person and man. These can be hierarchically linked as "A man is a person." Instances are concrete occurrences of abstract concepts. For example, we can have one concept, Man, with one instance of a Mike. Mike is a man and his first name is Mike. Axioms are rules that are valid in the modeled domain. There are simple symmetric, inverse, or transitive axioms consisting of several relations. For example, an inverse axiom is "If a person works for a company, the company employs this person."

Ontologies enable semantic interoperability between information systems, thereby serving a central role for the semantic Web and, in particular, serving as a means for the effective management of e-learning services. They can be used to specify user-oriented or domain-oriented learning services. Intelligent mediators can also use them: a central notion in teaching and learning. Therefore, the development of ontology can be useful for object or service modeling for e-learning domains.

There exist numerous scientific and commercial tools for the creation and maintenance of ontologies that have been used to build applications based on them, including those from the areas of knowledge management, engineering disciplines, medicine, and bioinformatics. It should be noted that ontologies do not overcome any interoperability problems per se since it is hardly conceivable that a single ontology is applied in all kinds of domains and applications. Ontology mapping does not intend to unify ontologies and their data, but to transform ontology instances according to the semantic relations defined at the conceptual level.

CONCLUSION

The semantic Web constitutes an environment in which human and machine agents will communicate on a semantic basis. This paper has examined current approaches used in managing learning objects. While it is clear that there is a comprehensive suite of standards that seem to have addressed some aspects of the management of learning objects, it is still clear that the management of interoperable learning objects is yet to be fully achieved. There are a lot of driving forces and a need for the development of flexible, portable, centralized, managed, and interoperable learning objects. Many challenges abound.

To meet these challenges, the author puts forward a new approach toward the management of interoperable learning objects by exploiting the power of ontologies and existing semantic and Web-services technology. It defines a framework that is being used toward enabling the semantic interoperability of learning services within the domain of e-learning. Further work is being done toward a definition of an ontology-management architecture for e-learning services. The architecture will define three main layers—interface, service integration, and management—with service composition running across all three. The aim of the architecture will be to provide an integration service platform that offers learner-centric support for Web-based learning, thus defining semantic relations between source learning resources (which may have been described using an ontology). This will be developed using Web services, an ontology, and agent components.

REFERENCES

Anderson, C. (2000). *eLearning: The definitions, the practice and the promise.* Carolina, USA: ICD Press.

Anido, L. E., Fernandez, M. J., Caeiro, M., Santos, J. M., Rodriguez, J. S., & Llamas, M. (2002). Educational metadata and brokerage for learning resources. *Computers and Education, 38*, 351-374.

Appelt, W. (1997, Spring). Basic support for cooperative work on the World Wide Web. International *Journal of Human Computer Studies: Special issue on Novel Applications of the WWW*. Cambridge: Academic Press.

Berners-Lee, T. (1999). *Weaving the Web* (pp. 35-43). San Francisco: Harper.

Berners-Lee, T., Henler, J., & Lassila, O. (2001). The semantic Web. *Scientific American, 5*.

Bonk, C. J., Kirkley, J. R., Hara, N., & Dennen, N. (2001). Finding the instructor in post-secondary online learning: Pedagogical, social, managerial, and technological locations. In J. Stephenson (Ed.), *Teaching and learning online: Pedagogies for new technologies* (pp. 76-97). London: Kogan Page.

Deek, F. P., Ho, K.-W., & Ramadhan, H. (2001). A review of Web-based learning systems for programming. *Proceedings of ED-MEDIA*, (pp. 382-387).

Erdmann, M., & Studer, R. (2000). How to structure and access XML documents with ontologies. *Data and Knowledge Engineering: Special Issues on Intelligent Information Integration DKE, 3*(36), 317-335.

Fensel, D., Horrocks, I., van Harmelen, F., McGuinness, D., & Patel-Schneider, P.F. (2001). OIL: Ontology infrastructure to enable the semantic Web. *IEEE Intelligent System, 16*(2).

Fetterman, D. M. (1998). Webs of meaning: Computer and Internet resources for educational research and instruction. *Educational Researcher, 27*(3), 22-30.

Gerhard, J., & Mayr, P. (2002). Competing in the e-learning environment: Strategies for universities. *35th Hawaii International Conference on Systems Sciences*, Big Island, HI.

GESTALK. (1998). *Getting educational systems to talk across leading edge technology project.* Retrieved August 2002 from *http://www.fdgroup.co.uk/gestalk*

Govindasmy, T. (2002). Successful implementation of e-learning pedagogical consideration. *The Internet and Higher Education, 4*, 287-289.

Gruber, T. R. (1995). Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies, 43*(5/6), 907-928.

Hamid, A. A. (2002). E-learning: Is it the "e" or the learning that matters? *The Internet and Higher Education, 4*, 287-289.

Harris, M. H. (1999). Is the revolution now over, or has it just begun? A year of the Internet in higher education. *The Internet & Higher Education, 1*(4), 243–251.

Horrocks, I. (2002). DAML+OIL: A description logic for semantic Web. *IEEE Bulletin of the Technical Committee on Data Engineering, 25*(1), 4-9.

Horrocks, I., & Tessaris, S. (2002). Querying the semantic Web: A formal approach. In I. Horrocks & J. Hendler (Eds.), *Proceedings of the 13th International Semantic Web Conference (ISWC 2002)*, *Lecture Notes in Computer Science*, 2342, (pp. 177-191). Springer-Verlag.

IEEE. (2002). *Learning Technologies Standardisation Committee.* Retrieved August 2002 from *http://ltsc.ieee.org*

Ishaya, T., Jenkins, C., & Goussios, S. (2002). The role of multimedia and software agents for effective online learning. *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, (pp. 135-138).

Jack, A., Bonk, C. J., & Jacobs, F. R. (2002). Twenty-first century college syllabi: Options for online communication and interactivity. *The Internet and Higher Education, 5*(1), 1-19.

Jennings, N. R. (2000). On agent-based software engineering. *Artificial Intelligence, 117*(2), 277-296.

Jesshope, C. (1999). Web-based teaching: Tools and experiences. *Austrian Computer Science Communication, 21*(1), 27-38.

Jesshope, C., Heinrich, E., & Kinshuk. (2000). Technology integrated learning environments for education at a distance. *DEANZ 2000 Conference*, 26-29.

Koper, R. (2001). *Modelling units of study from a pedagogical perspective: The pedagogical metmodel behind EML.* Retrieved August 2002 from *http://eml.ou.nl/introduction/articles.htm*

Koper, R. (2002). *Educational modelling language: Adding instructional design to existing specification.* Retrieved August 2002 from *http://wwrz.uni-frankfurt.de*

Lim, W. (1998). *Managing software reuse: A comprehensive guide to strategically*

reengineering the organisation for reusable components. Upper Saddle River, NJ: Prentice-Hall.

MacColl, J. (2001). Virtuous learning environments: The library and the VLE. *Program, 35*(3), 227-239.

Manouselis, N., Panagiotou, K., Psichidou, R., & Sampson, D. (2002). Issues in designing Web-based environment for learning communities with special educational needs. *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, (pp. 239-243).

Mergendoller, J. R. (1996). Moving from technological possibility to richer student learning: Revitalized infrastructure and reconstructed pedagogy. *Educational Researcher, 25*(8), 43-46.

PROMETEUS. (2002). *Promoting multimedia access to education and training in European society.* Retrieved August 2002 from *http://prometeus.org*

Wooldridge, M. (1997). Agent-based software engineering. *IEEE Proceedings in Software Engineering, 144*(1), 26-37.

Yazon, J. M. O., Mayer-Smith, J. A., & Redfield, R. J. (2002). Does the medium change the message? The impact of a Web-based genetics course on university students' perspectives on learning and teaching. *Computers & Education, 38*(1-3), 267-285.

KEY TERMS

Electronic Learning (E-Learning): Defined as a special kind of technology-based learning. E-learning systems and tools bring geographically dispersed teams together for learning across great distances. It is now one of the fastest growing trends in computing and higher education.

Interoperability: Ability to work together, sharing information, capabilities, or other specific goals while being different at some technological level.

Learning-Object Metadata (LOM): Metadata that contain semantic information about learning objects. The main aim of LOM specification is to enable the reuse, search, and retrieval of learning objects. The standard, developed by the IEEE Learning Technology Standards Committee (LTSC) in 1997, specifies a conceptual data schema that defines the structure of metadata instances for a learning object.

Learning Objects: Defined as any entity—digital or nondigital—that may be used, reused, or referenced for learning, education, or training. Examples of learning objects include multimedia content, instructional content, learning objectives, instructional software and software tools, people, organizations, and events referenced during technology-supported learning.

Ontologies: An ontology is an explicit, formal specification of a shared conceptualization of a domain of interest. This indicates that an onotology should be machine readable (which excludes natural language). It indicates that it captures consensual knowledge that is not private to an individual, but accepted as a group or committee of practice. The reference to a domain of interest indicates that domain ontologies do not model the whole world, but rather model just parts that are relevant to the task at hand.

Resource Description Framework (RDF): RDF provides means for adding semantics to a document. It is an infrastructure that enables the encoding, exchange, and reuse of information-structured metadata. RDF allows multiple metadata schemas to be read by humans as well as machines, providing interoperability between applications that exchange machine-understandable information on the Web.

Semantic Web: The semantic Web constitutes an environment in which human and machine agents will communicate on a semantic basis. It is to be achieved via semantic markup and metadata annotations that describes content and functions.

Shareable Content Object Reference Model (SCORM): SCORM is an XML-based framework used to define and access information about learning objects so they can be easily shared among different learning-management systems. The SCORM specifications, which are distributed through the Advanced Distributed Learning (ADL) Initiative Network, define an XML-based means of representing course structures, an application programming interface, a content-to-LMS data model, a content launch specification, and a specification for metadata information for all components of a system.

Intrusion Detection Systems

H. Gunes Kayacik
Dalhousie University, Canada

A. Nur Zincir-Heywood
Dalhousie University, Canada

Malcolm I. Heywood
Dalhousie University, Canada

INTRODUCTION

Along with its numerous benefits, the Internet also created numerous ways to compromise the security and stability of the systems connected to it. In 2003, 137529 incidents were reported to CERT/CC© while in 1999, there were 9859 reported incidents (CERT/CC©, 2003). Operations, which are primarily designed to protect the availability, confidentiality, and integrity of critical network information systems, are considered to be within the scope of security management. Security management operations protect computer networks against denial-of-service attacks, unauthorized disclosure of information, and the modification or destruction of data. Moreover, the automated detection and immediate reporting of these events are required in order to provide the basis for a timely response to attacks (Bass, 2000). Security management plays an important, albeit often neglected, role in network management tasks.

Defensive operations can be categorized in two groups: static and dynamic. Static defense mechanisms are analogous to the fences around the premises of a building. In other words, static defensive operations are intended to provide barriers to attacks. Keeping operating systems and other software up-to-date and deploying firewalls at entry points are examples of static defense solutions. Frequent software updates can remove software vulnerabilities that are susceptible to exploits. By providing access control at entry points, they therefore function in much the same way as a physical gate on a house. In other words, the objective of a firewall is to keep intruders out rather than catching them. Static defense mechanisms are the first line of defense, they are relatively easy to deploy and naturally provide significant defense improvement compared to the initial unguarded state of the computer network. Moreover they act as the foundation for more sophisticated defense mechanisms.

No system is totally foolproof. It is safe to assume that intruders are always one step ahead in finding security holes in current systems. This calls attention to the need for dynamic defenses. Dynamic defense mechanisms are analogous to burglar alarms, which monitor the premises to find evidence of break-ins. Built upon static defense mechanisms, dynamic defense operations aim to catch the attacks and log information about the incidents such as source and nature of the attack. Therefore, dynamic defense operations accompany the static defense operations to provide comprehensive information about the state of the computer networks and connected systems.

Intrusion detection systems are examples of dynamic defense mechanisms. An intrusion detection system (IDS) is a combination of software and hardware, which collects and analyzes data collected from networks and the connected systems to determine if there is an attack (Allen, Christie, Fithen, McHugh, Pickel, & Stoner, 1999). Intrusion detection systems complement static defense mechanisms by double-checking firewall configuration, and then attempt to catch attacks that firewalls let in or never perceive (such as insider attacks). IDSs are generally analyzed from two aspects:

- **IDS Deployment:** Whether to monitor incoming traffic or host information.
- **Detection Methodologies:** Whether to employ the signatures of known attacks or to employ the models of normal behavior.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Regardless of the aspects above, intrusion detection systems correspond to today's dynamic defense mechanisms. Although they are not flawless, current intrusion detection systems are an essential part of the formulation of an entire defense policy.

DETECTION METHODOLOGIES

Different detection methodologies can be employed to search for the evidence of attacks. Two major categories exist as detection methodologies: misuse and anomaly detection. Misuse detection systems rely on the definitions of misuse patterns i.e., the descriptions of attacks or unauthorized actions (Kemmerer & Vigna, 2002). A misuse pattern should summarize the distinctive features of an attack and is often called the signature of the attack in question. In the case of signature based IDS, when a signature appears on the resource monitored, the IDS records the relevant information about the incident in a log file. Signature based systems are the most common examples of misuse detection systems. In terms of advantages, signature based systems, by definition, are very accurate at detecting known attacks, where these are detailed in their signature database. Moreover, since signatures are associated with specific misuse behavior, it is easy to determine the attack type. On the other hand, their detection capabilities are limited to those within signature database. As new attacks are discovered, a signature database requires continuous updating to include the new attack signatures, resulting in potential scalability problems.

As opposed to misuse IDSs, anomaly detection systems utilize models of the acceptable behavior of the users. These models are also referred to as normal behavior models. Anomaly based IDSs search for any deviation from the (characterized) normal behavior. Deviations from the normal behavior are considered as anomalies or attacks. As an advantage over signature based systems, anomaly based systems can detect known and unknown (i.e., new) attacks as long as the attack behavior deviates sufficiently from the normal behavior. However, if the attack is similar to the normal behavior, it may not be detected. Moreover, it is difficult to associate deviations with specific attacks since the anomaly based IDSs only utilize models of normal behavior. As the users change their behavior as a result of additional service or hardware, even the normal activities of a user may start raising alarms. In that case, models of normal behavior require be redefinition in order to maintain the effectiveness of the anomaly based IDS.

Human input is essential to maintain the accuracy of the system. In the case of signature based systems, as new attacks are discovered, security experts examine the attacks to create corresponding detection signatures. In the case of anomaly systems, experts are needed to define the normal behavior. Therefore, regardless of the detection methodology, frequent maintenance is essential to uphold the performance of the IDS.

Given the importance of IDSs, It is imperative to test them to determine their performance and eliminate their weaknesses. For this purpose, researchers conduct tests on standard benchmarks (Kayacik, Zincir, & Heywood, 2003; Pickering, 2002). When measuring the performance of intrusion detection systems, the detection and false positive rates are used to summarize different characteristics of classification accuracy. In simple terms, false positives (or false alarms) are the alarms generated by a non-existent attack. For instance, if an IDS raises alarms for the legitimate activity of a user, these log entries are false alarms. On the other hand, detection rate is the number of correctly identified attacks over all attack instances, where correct identification implies the attack is detected by its distinctive features. An intrusion detection system becomes more accurate as it detects more attacks and raises fewer false alarms. A receiver operating characteristic or ROC, where this details how system performance varies as a function of different parameters, typically characterizes the sensitivity of the IDS.

IDS DEPLOYMENT STRATEGIES

In addition to the detection methodologies, data is collected from two main sources: traffic passing through the network and the hosts connected to the network. Therefore, according to where they are deployed, IDSs are divided into two categories, those that analyze network traffic and those that analyze information available on hosts such as operating system audit trails. The current trend in intrusion detection is to combine both host based and network based information to develop hybrid systems and

therefore not rely on any one methodology. In both approaches however, the amount of audit data is extensive, thus incurring large processing overheads. A balance therefore exists between the use of resources, and the accuracy and timeliness of intrusion detection information.

Network based IDS inspect the packets passing through the network for signs of an attack. However, the amount of data passing through the network stream is extensive, resulting in a trade off between the number of detectors and the amount of analysis each detector performs. Depending on throughput requirements, a network based IDS may inspect only packet headers or include the content. Moreover, multiple detectors are typically employed at strategic locations in order to distribute the task. Conversely, when deploying attacks, intruders can evade IDSs by altering the traffic. For instance, fragmenting the content into smaller packets causes IDSs to see one piece of the attack data at a time, which is insufficient to detect the attack. Thus, network based IDSs, which perform content inspection, need to assemble the received packets and maintain state information of the open connections, where this becomes increasingly difficult if a detector only receives part of the original attack or becomes "flooded" with packets.

A host-based IDS monitors resources such as system logs, file systems, processor, and disk resources. Example signs of intrusion on host resources are critical file modifications, segmentation fault errors, crashed services, or extensive usage of the processors. As opposed to network-based IDSs, host-based IDSs can detect attacks that are transmitted over an encrypted channel. Moreover, information regarding the software that is running on the host is available to host-based IDS. For instance, an attack targeting an exploit on an older version of a Web server might be harmless for the recent versions. Network-based IDSs have no way of determining whether the exploit has a success chance, or of using a priori information to constrain the database of potential attacks. Moreover, network management practices are often critical in simplifying the IDS problem by providing appropriate behavioral constraints, thus making it significantly more difficult to hide malicious behaviors (Cunningham, Lippmann, & Webster, 2001).

CHALLENGES

The intrusion detection problem has three basic competing requirements: speed, accuracy, and adaptability. The speed problem represents a quality of service issue. The more analysis (accurate) the detector, the higher the computational overhead. Conversely, accuracy requires sufficient time and information to provide a useful detector. Moreover, the rapid introduction of both new exploits and the corresponding rate of propagation require that detectors be based on a very flexible/scalable architecture. In today's network technology where gigabit Ethernet is widely available, existing systems face significant challenges merely to maintain pace with current data streams (Kemmerer & Vigna, 2002).

An intrusion detection system becomes more accurate as it detects more attacks and raises fewer false alarms. IDSs that monitor highly active resources are likely to have large logs, which in turn complicate the analysis. If such an IDS has a high false alarm rate, the administrator will have to sift through thousands of log entries, which actually represent normal events, to find the attack-related entries. Therefore, increasing false alarm rates will decrease the administrator's confidence in the IDS. Moreover, intrusion detection systems are still reliant on human input in order to maintain the accuracy of the system. In case of signature based systems, as new attacks are discovered, security experts examine the attacks to create corresponding detection signatures. In the case of anomaly systems, experts are needed to define the normal behavior. This leads to the adaptability problem. The capability of the current intrusion detection systems for adaptation is very limited. This makes them inefficient in detecting new or unknown attacks or adapting to changing environments (i.e., human intervention is always required). Although a new research area, incorporation of machine learning algorithms provides a potential solution for accuracy and adaptability of the intrusion detection problem.

CURRENT EXAMPLES OF IDS

Intrusion detection systems reviewed here are by no means a complete list but a subset of open source and commercial products, which are intended to provide readers different intrusion detection practices.

- **Snort:** Snort is one of the best-known lightweight IDSs, which focuses on performance, flexibility, and simplicity. It is an open-source intrusion detection system that is now in quite widespread use (Roesch, 1999). Snort is a network based IDS which employs signature based detection methods. It can detect various attacks and Probes including instances of buffer overflows, stealth port scans, common gateway interface attacks, and service message block system Probes (Roesch, 1999). Hence, Snort is an example of active intrusion detection systems that detects possible attacks or access violations while they are occurring (CERT/CC ©, 2001).
- **Cisco IOS (IDS Component):** Cisco IOS provides a cost effective way to deploy a firewall with network based intrusion detection capabilities. In addition to the firewall features, Cisco IOS Firewall has 59 built-in, static signatures to detect common attacks and misuse attempts (Cisco Systems, 2003). The IDS process on the firewall router inspects packet headers for intrusion detection by using those 59 signatures. In some cases routers may examine the whole packet and maintain the state information for the connection. Upon attack detection, the firewall can be configured to log the incident, drop the packet, or reset the connection.
- **Tripwire:** When an attack takes place, attackers usually replace critical system files with their versions to inflict damage. Tripwire (Tripwire Web Site, 2004) is an open-source host-based tool, which performs periodic checks to determine which files are modified in the file system. To do so, Tripwire takes snapshots of critical files. Snapshot is a unique mathematical signature of the file where even the smallest change results in a different snapshot. If the file is modified, the new snapshot will be different than the old one; therefore critical file modification would be detected. Tripwire is different from the other intrusion detection systems because rather than looking for signs of intrusion, Tripwire looks for file modifications.

FUTURE TRENDS

As indicated above, various machine learning approaches have been proposed in an attempt to improve on the generic signature-based IDS. The basic motivation is to measure how close a behavior is to some previously established gold standard of misuse or normal behavior. Depending on the level of a priori or domain knowledge, it may be possible to design detectors for specific categories of attack (e.g., denial of service, user to root, remote to local). Generic machine learning approaches include clustering or data mining in which case the data is effectively unlabeled. The overriding assumption is that behaviors are sufficiently different for normal and abnormal behaviors to fall into different "clusters". Specific examples of such algorithms include artificial immune systems (Hofmeyr & Forrest, 2000) as well as various neural network (Kayacik, Zincir-Heywood, & Heywood, 2003; Lee & Heinbuch, 2001) and clustering algorithms (Eskin, Arnold, Prerau, Portnoy, & Stolfo, 2002).

Naturally the usefulness of machine learning systems is influenced by the features on which the approach is based (Lee & Stolfo, 2001). Domain knowledge that has the capability to significantly simplify detectors utilizing machine learning often make use of the fact that attacks are specific to protocol-service combinations. Thus, first partitioning data based on the protocol-service combination significantly simplifies the task of the detector (Ramadas, Ostermann, & Tjaden, 2003).

When labeled data is available then supervised learning algorithms are more appropriate. Again any number of machine learning approaches have been proposed, including: decision trees (Elkan, 2000), neural networks (Hofmann & Sick, 2003) and genetic programming (Song, Heywood, & Zincir-Heywood, 2003). However, irrespective of the particular machine learning methodology, all such methods need to address the scalability problem. That is to say, datasets characterizing the IDS problem are exceptionally large (by machine learning standards). Moreover, the continuing evolution of the base of attacks also requires that any machine learning approach also have

the capability for online or incremental learning. Finally, to be of use to network management practitioners it would also be useful if machine learning solutions were transparent. That is to say, rather than provide "black box solutions", it is much more desirable if solutions could be reverse engineered for verification purposes. Many of these issues are still outstanding. For example cases that explicitly provide scalable solutions (Song, Heywood, & Zincir-Heywood, 2003) or automatically identify weaknesses in the IDS (Dozier, Brown, Cain, & Hurley, 2004) are only just appearing.

CONCLUSION

An intrusion detection system is a crucial part of the defensive operations that complements the static defenses such as firewalls. Essentially, intrusion detection systems search for signs of an attack and flag when an intrusion is detected. In some cases they may take an action to stop the attack by closing the connection or report the incident for further analysis by network administrators. According to the detection methodology, intrusion detection systems are typically categorized as misuse detection and anomaly detection systems. From a deployment perspective, they are be classified as network based or host based although such distinction is coming to an end in today's intrusion detection systems where information is collected from both network and host resources. In terms of performance, an intrusion detection system becomes more accurate as it detects more attacks and raises fewer false alarms. Future advances in IDS are likely to continue to integrate more information from multiple sources (sensor fusion) whilst making further use of artificial intelligence to minimize the size of log files necessary to support signature databases. Human intervention, however, is certainly necessary and set to continue for the foreseeable future.

REFERENCES

Allen, J., Christie, A., Fithen, W., McHugh, J., Pickel, J., & Stoner, E. (1999). *State of the practice of intrusion detection technologies.* CMU/SEI Technical Report (CMU/SEI-99-TR-028). Retrieved June 2004 from *http://www.sei.cmu.edu/publications/documents/99.reports/99tr028/99tr028abstract.html*

Bass, T. (2000). Intrusion detection systems and multisensor data fusion, *Communications of the ACM, 43*(4), 99-105.

CERT/CC© (2001). Identifying tools that aid in detecting signs of intrusion. Retrieved from *http://www.cert.org/security-improvement/implementations/i042.07.html*

CERT/CC© (2003). Incident Statistics 1988-2003. Retrieved June 2004 from *http://www.cert.org/stats/*

Cisco Systems Inc. (2003). Cisco IOS Firewall Intrusion Detection System Documentation. Retrieved June 2004 from *http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newft/120t/120t5/iosfw2/ios_ids.htm*

Cunningham, R.K., Lippmann, R.P., & Webster S.E. (2001). Detecting and displaying novel computer attacks with macroscope. *IEEE Transactions on Systems, Man, and Cybernetics – Part A, 31*(4), 275-280.

Dozier, G., Brown, D., Cain, K., & Hurley, J. (2004). Vulnerability analysis of immunity-based intrusion detection systems using evolutionary hackers. *Proceedings of the Genetic and Evolutionary Computation Conference, Lecture Notes in Computer Science, 3102*, (pp. 263-274).

Elkan, C. (2000). Results of the KDD'99 classifier learning. *ACM SIGKDD Explorations, 1*, 63-64.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection: Detecting attacks in unlabeled data. In D. Barbara & S. Jajodia (Eds.), *Applications of data mining in computer security*. Kluwer Academic.

Hofman, A. & Sick, B. (2003). Evolutionary optimization of radial basis function networks for intrusion detection. *Proceedings of the International Joint IEEE-INNS Conference on Neural Networks*, (pp. 415-420).

Hofmeyr, S.A. & Forrest, S. (2000). Architecture for an Artificial Immune System. *Evolutionary Computation, 8*(4), 443-473.

Kayacik, G. & Zincir-Heywood, N. (2003). A case study of three open source security management tools. *Proceedings of 8th IFIP/IEEE International Symposium on Integrated Network Management*, (pp. 101-104).

Kayacik, G., Zincir-Heywood, N., & Heywood, M. (2003). On the capability of an SOM based intrusion detection system. *Proceedings of the International Joint IEEE-INNS Conference on Neural Networks*, (pp. 1808-1813).

Kemmerer, R.A. & Vigna, G. (2002). Intrusion detection: A brief history and overview. *IEEE Security and Privacy*, 27-29.

Lee, S.C. & Heinhuch, D.V. (2001). Training a neural-network based intrusion detector to recognize novel attacks. *IEEE Transactions on Systems, Man, and Cybernetics – Part A, 31*(4), 294-299.

Pickering, K. (2002). *Evaluating the viability of intrusion detection system benchmarking.* BS Thesis submitted to The Faculty of the School of Engineering and Applied Science, University of Virginia. Retrieved June 2004 from *http://www.cs.virginia.edu/~evans/students.html*

Ramadas, M., Ostermann, S., & Tjaden, B. (2003). Detecting anomalous network traffic with self-organizing maps. *The 6th International Symposium on Recent Advances in Intrusion Detection, Lecture Notes in Computer Science, 2820*, (pp. 36-54).

Roesch, M. (1999). Snort – Lightweight intrusion detection for networks. *Proceedings of the 13th Systems Administration Conference*, (pp. 229-238).

Song, D., Heywood, M.I., & Zincir-Heywood, A.N. (2003). A linear genetic programming approach to intrusion detection. *Proceedings of the Genetic and Evolutionary Computation Conference. Lecture Notes in Computer Science, 2724*, (pp. 2325-2336).

Tripwire Web Site (2004). Home of the Tripwire Open Source Project. Retrieved June 2004 from *http://www.tripwire.org/*

KEY TERMS

Attack vs. Intrusion: A subtle difference, intrusions are the attacks that succeed. Therefore, the term attack represents both successful and attempted intrusions.

CERT/CC©: CERT Coordination Center. Computer security incident response team, which provide technical assistance, analyze the trends of attacks and provide response for incidents. Documentation and statistics are published at their Web site www.cert.org.

Exploit: Taking advantage of a software vulnerability to carry out an attack. To minimize the risk of exploits, security updates, or software patches should be applied frequently.

Fragmentation: When the data packet is too large to transfer on given network, it is divided into smaller packets. These smaller packets are reassembled on destination host. Among with other methods, intruders can deliberately divide the data packets to evade IDSs.

Light Weight IDS: An intrusion detection system, which is easy to deploy and has a small footprint on system resources.

Logging: Recording vital information about an incident. Recorded information should be sufficient to identify the time, origin, target, and if applicable characteristics of the attack.

Machine Learning: A research area of artificial intelligence, which is interested in developing solutions from data or in interactive environment alone.

Open Source Software: Software with its source code available for users to inspect and modify to build different versions.

Security Management: In network management, the task of defining and enforcing rules and regulations regarding the use of the resources.

Investment Strategy for Integrating Wireless Technology into Organizations

Assion Lawson-Body
University of North Dakota, USA

INTRODUCTION

Firms rely on IT investments (Demirhan et al., 2002; Tuten, 2003), because a growing number of executives believe that investments in information technology (IT) (i.e., wireless technologies) help boost firm performance. The use of wireless communications and computing is growing quickly (Kim & Steinfield, 2004; Leung & Cheung, 2004; Yang et al., 2004). But issues of risk and uncertainty due to technical, organizational, and environmental factors continue to hinder executive efforts to produce meaningful evaluation of investment in wireless technology (Smith et al., 2002). Despite the use of investment appraisal techniques, executives often are forced to rely on instinct when finalizing wireless investment decisions. A key problem with evaluation techniques, it emerges, is their treatment of uncertainty and their failure to account for the fact that outside of a decision to reject an investment outright, firms may have an option to defer an investment until a later period (Tallon et al., 2002).

Utilization of wireless devices and being connected without wires is inevitable (Gebauer et al., 2004; Jarvenpaa et al, 2003). Market researchers predict that by the end of 2005, there will be almost 500 million users of wireless devices, generating more than $200 billion in revenues (Chang & Kannan, 2002; Xin, 2004). By 2006, the global mobile commerce (m-commerce) market will be worth $230 billion (Chang & Kannan, 2002). Such predictions indicate the importance that is attached to wireless technologies as a way of supporting business activities. Evaluating investments in wireless technology and understanding which technology makes the best fit for a company or organization is difficult because of the numerous technologies and the costs, risks, and potential benefits associated with each technology.

The purpose of this study is twofold: first, to identify and discuss different investment options; and second, to assist in formulating an investment strategy for integrating wireless technologies into organizations.

This article is organized as follows: Section II contains major uncertainties and risks in the field of wireless technologies. In Section III, wireless technology and IT investment tools are examined. In Section IV, formulating a wireless technology investment strategy is discussed. The conclusion of this article is presented in Section V.

MAJOR UNCERTAINTIES AND RISKS IN THE FIELD OF WIRELESS TECHNOLOGIES

Businesses today face several uncertainties in effectively using wireless technology (Shim et al., 2003; Yang et al., 2004). One of the first uncertainties for managers investing in wireless technology is that standards may vary from country to country, making it difficult for devices to interface with networks in different locations (Shim et al., 2003; Tarasewich et al., 2002).

Another uncertainty is that wireless networks lack the bandwidth of their wired counterparts (Tarasewich et al., 2002). Applications that run well on a wired network may encounter new problems with data availability, processing efficiency, concurrency control, and fault tolerance when ported to a mobile environment. Limited bandwidth inhibits the amount and types of data that can be transmitted to mobile devices. Significantly improved bandwidth is clearly needed before new types of mobile applications such as Web access, video, document transfer, and database access can be implemented. Bandwidth is expected to increase rapidly over the next few years with the introduction of a new generation of wireless technologies. It is uncertain, therefore, how fast firms will follow the increased bandwidth evolution.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

User interface is another uncertainty related to the development of wireless technology (Shim et al., 2003). Mobile devices provide very restrictive user interfaces that limit possible employee and consumer uses of mobile technology. The ideal mobile user interface will exploit multiple input/output technologies. The employee should be able to switch effortlessly from text-based screens to streaming audio/video to voice-powered interaction. Mobile users require different input and output methods in different situations. It is necessary to create a range of standard interfaces that can be reused in different mobile devices. As wireless technology development promises to improve this interface with such features as voice recognition, voice synthesis, and flexible screens, increased usage likely will result. New and more powerful user interfaces are essential to 3G (three-generation) wireless success. Finally, security is another uncertainty related to wireless technologies (Shim et al., 2003).

Where uncertainties exist, they are viewed as risks that will reduce the potential payoff of investment in wireless technology. Thus, organizations may be hesitant to invest in a particular technology, because they are afraid of high costs associated with potential obsolescence of technologies in which they may have invested.

Given all these uncertainties and risks, past research on IT investments should be analyzed to provide a basis for understanding investment in wireless technology.

WIRELESS TECHNOLOGY AND INFORMATION TECHNOLOGY INVESTMENT TOOLS

IT investment justification models can vary from intuition-based cost-benefit analysis, regression analysis, payback rules, accounting rates of return, and financial and economic models such as Net Present Value (NPV), to Real Options analysis (ROA) (Kohli & Sherer, 2002; Walters & Giles, 2000).

Cost-Benefit Analysis

Cost-benefit analysis often requires substantial data collection and analysis of a variety of costs and benefits. However, most IT investments and their benefits involve great complexity and require a detailed cost-benefit analysis. This analysis involves explicitly spelling out the costs and benefits in a formula such as an equation for an investment that improves productivity (Kohli & Sherer, 2002).

Regression Analysis

Some authors use statistical analysis (e.g., regression analysis) to understand the relationship between the IT investment and payoff. They usually examine the correlation table, listing the strength of relationship between the investment (independent) variables, and the payoff (dependent) variables.

Payback Rules

Payback rules track how many periods IT managers must wait before cumulated cash flows from the project exceed the cost of the investment project (Walters & Giles, 2000). If this number of periods is less than or equal to the firm's benchmark, the project gets the go-ahead (Walters & Giles, 2000).

Accounting Rates of Return

An accounting rate of return is the ratio of the average forecast profits over the project's lifetime (after depreciation and tax) to the average book value of the IT investment (Walters & Giles, 2000). Again, comparison with a threshold rate is sought before investment goes ahead (Walters & Giles, 2000).

Payback rules and accounting rates of return do not take into account uncertainties and risks. Therefore, they are not adequate to analyze investment strategy in wireless technologies.

Net Present Value (NPV) Analysis

The time value of investment is represented in NPV. The NPV rule assumes that either the investment is reversible, or, if the investment is irreversible, the firm can only invest now, otherwise it will never be able to do so in the future (Tallon et al., 2002). While NPV provides information about the time value of the investment, it does not take into account the risks or opportunities created by stopping, decreasing, or increasing investment in the future (Kohli & Sherer, 2002). In fact, the NPV has been criticized widely

because of its inability to model uncertainty, a factor that is particularly relevant in the context of IT investment decisions (Tallon et al., 2002; Tuten, 2003). In evaluating IT investments that exhibit high growth potential and high uncertainty, NPV is inadequate, but Real Options Analysis (ROA) seems to be a better tool (Tallon et al., 2002).

Using an NPV method of wireless technology investment analysis means that once the decision is made not to invest because of security issues, bandwidth limitations, or standard issues, it likely will not be revisited for some time. The NPV method will allow managers to mismanage investments in wireless technologies, because a firm could invest in a wireless technology with high cost and an uncertain payoff. With NPV, it is difficult to obtain accurate estimates of revenues and costs. In the absence of accurate estimates, NPV may lead to an erroneous decision.

Real Options Analysis (ROA)

The Real Options approach helps managers understand the potential payoff from IT investments in a multi-phase investment scenario (Kohli & Sherer, 2002). Real option theory recognizes that the ability to delay, suspend, or abandon a project is valuable when the merits of the project are uncertain (Tallon et al., 2002).

In practice, the application of Real Options has been proven difficult, though not impossible (Tallon et al., 2002). ROA remains a controversial technique because it is based on decision tree analysis, which tends to include too much detail in the cash flow portion of the model. ROA in practice is inherently complex, because there are many assumptions behind the different models used with it. In the context of wireless technology, some of those assumptions may be questionable. For example, few executives could assign a credible market value to an IT investment, especially where it is part of a multi-phase project, such as upgrading network capacity as part of a wireless networking strategy (Tallon et al., 2002).

Despite any initial misgivings, the benefits of ROA remain attractive to wireless technology managers, who are repeatedly faced with difficult investment decisions involving technical and organizational uncertainty, multiple forms of risk, and incomplete information. ROA is a positive step because it allows wireless technology implementation decision makers to consider risk and uncertainty factors in their investment decisions (Tallon et al., 2002).

ROA may be used to evaluate investment in wireless technology. An option gives the holder the right to invest now or at a future point in time (Tallon et al., 2002). If future developments of wireless

Table 1. NPV vs. ROA

NPV	ROA
Managers are passive investors.	Managers are active investors.
Managers do not have the flexibility to sell the asset.	Managers have the flexibility to sell the asset.
An NPV calculation only uses information that is known at the time of the appraisal. The choice is all-or-nothing.	An ROA uses initial choice followed by more choices as information becomes available.
NPV does not take into account uncertainties and risks.	ROA takes into account uncertainties and risks.
Managers do not have the flexibility to invest further, wait and see, or abandon the project entirely.	Managers have the flexibility to invest further, wait and see, or abandon the project entirely.
According to NPV theory, the future cash flows of an investment project are estimated.	By contrast, real options calculations involve a wide range of future cash flow probability distributions.
NPV does not use a decision tree analysis.	Real options theory is related to decision tree analysis.
With NPV, subsequent decisions cannot modify the project once it is undertaken.	With Real options theory, subsequent decisions can modify the project once it is undertaken.

technologies remove or otherwise reduce a key source of uncertainty to some satisfactory level, the firm may exercise its option and proceed with a full-blown implementation of the wireless technology investment. If, however, the uncertainty continues or is not adequately resolved, the expiration period can be extended, thus reducing any risk of future losses. In high-risk areas involving emerging technologies such as wireless telecommunications, ROA is useful for discovering investment possibilities, particularly for firms seeking to acquire a first-mover advantage (Kulatilaka & Venkatraman, 2001). With ROA, firms may consider even an initial investment or small-scale pilot investment (Tallon et al., 2002).

NPV vs. ROA

Table 1 shows the comparison between NPV and ROA, because both are the most used by IT and wireless technology executives (Tallon et al., 2002).

FORMULATING A WIRELESS TECHNOLOGY INVESTMENT STRATEGY

We argue that the use of Real Option technique is appropriate to analyze investment in wireless technologies. Also, the amount or level of commitment an organization will take, related to purchasing and implementing a given technology (in this case, wireless technologies) must be considered. The different options of investment that exist in accordance with the amount or level of commitment of an organization are the following (Smith et al., 2002):

- Growth option
- Staging option
- Exit option
- Sourcing option
- Business scope option
- Learning option

Growth option is an investment choice that would allow a company to invest with the intention that the expenditure could produce opportunities for the company that would be more beneficial than just the initial benefits produced by the technology. These opportunities can occur anywhere from immediately to well after the technology is implemented. Being able to provide an additional service that becomes lucrative and stems from the technology would be an example of a growth option.

Financing a technology or purchasing a technology in parts or stages is considered the staging option. The benefit of purchasing a wireless technology in stages is that it allows managers to make decisions before each additional purchase or stage. The benefits can be reevaluated before each additional expenditure to see if there is any marginal benefit and if further investment is needed. In wireless technology, such as wireless LANs (WLANs), an initial access point (AP) can be set up, and, if successful, further APs can be purchased and implemented.

If there is a current activity conducted by the business or organization that is not producing any clear benefit and is a high expense, the business or organization may want to slowly taper off such an activity. It may do this through the purchase of a standard technology. Then, the firm would be able to outsource to, partner with, or align with another organization to handle such an activity. This is classified as an exit option. An example of this could be a coffee shop that provides patrons with wired Internet access while they are at the coffee shop. If the cost of running a server with wired capabilities for Internet access became too costly, then the coffee shop could purchase an AP with 802.11b (a common standard in WLANs) to provide Internet access to patrons that have wireless network interface cards (NICs). The coffee shop could be doing this before they outsource to or partner with another company that can provide such a service more cheaply than the coffee shop itself could maintain.

Sourcing options occur when a company or organization chooses to invest in a technology for the purpose of adding input sources, channels, and/or platforms (Smith et al., 2002). A wireless example of this is a firm purchasing a printer that allows for Bluetooth and/or infrared communications in order to provide the advantage of accepting multiple inputs for the printer instead of inputting solely through a universal serial bus (USB) or parallel port. This would allow for printing from handheld devices (i.e., a different type of device besides notebooks or desktops).

Another option is a business scope option. This option provides a firm with the ability to "add to or adapt the product/service mix of the firm quickly and efficiently" (Smith et al., 2002). Using the coffee shop example again, a coffee shop that has no current offerings of Internet service to its patrons could add wireless APs to provide its patrons with Internet service, thus adding to the services the coffee shop provides.

The last option—learning—is when a company invests primarily for the experience of gaining more knowledge about a new technology. A technology consulting firm would invest perhaps in new WLAN technologies in order to fully test and learn such technologies so that managers could then recommend these technologies to customers and fully explain these technologies to them, as well.

From our earlier discussion, we find that with an ROA, unlike with an NPV, a firm with an investment opportunity has an option to invest now or in the future. Once the company exercises its option to invest, the lost option value is part of the opportunity cost of the investment. Our study demonstrates that when there is high uncertainty, the option value of an investment is significant. Therefore, because of the innovative nature of wireless technology, it is preferable to use the combined approach of Real Option and the amount or level of commitment an organization will take, related to investing in wireless technologies to evaluate uncertain wireless technology implementation projects. We formulate that both of these ideas can coexist. Depending on the way a firm or organization chooses to implement the wireless technology, a different impact can be expected.

CONCLUSION

The objective of this article is to examine different wireless technology investment tools and formulate an appropriate wireless technology investment strategy.

This research begins with the presentation of several uncertainties and risks in the field of wireless technology. First, the field has no single and universally accepted standard. Wireless networks lack the bandwidth of their wired counterparts. User interface is an uncertainty related to the development of wireless technology. Finally security is another uncertainty related to wireless technologies.

Since uncertainties and risks exist in the wireless technology field, organizations planning to invest in wireless technology implementation have to use an investment analysis tool that takes those uncertainties and risks into account. In order to identify the investment strategy that fits with wireless technology, this article has analyzed the different investment tools such as cost-benefit analysis, regression analysis, payback rules, accounting rates of return, NPV, and ROA.

For multi-period investment decisions, ROA is superior to other investment tools and the ubiquitous net present value (NPV) approach. In a world of uncertainty such as a wireless technology implementation project, real options offer the flexibility to expand, extend, contract, abandon, or defer a project in response to unforeseen events that drive the value of a project up or down over time.

The main contribution of this study is the formulation of an appropriate wireless technology investment strategy.This study recommends the combined use of ROA and the level of commitment of an organization. The different options for investment (growth option, staging option, exit option, sourcing option, business scope option, and learning option), which exist in accordance with the amount or level of commitment of an organization, were presented, discussed, and illustrated in relation to wireless technology.

Clearly, the concept of a combined approach developed in this research, based on the use of real option and the level of commitment of an organization, offers much promise for future study. Therefore, we encourage our IS colleagues to accept the challenges that the objective of this article posed. Future research is necessary, because wireless technology evolves so rapidly. Additional research also should expand the range of the IT investment tools and examine their effects on the decision to invest in wireless technology implementation.

REFERENCES

Chang, A., & Kannan, P. (2002). Preparing for wireless and mobile technologies in government. *E-Government Series*, 1-42.

Demirhan, D., Jacob, V., & Raghunathan, S. (2002). Strategic IT investments: Impacts of switching cost and declining technology cost. *Proceedings of the 23rd International Conference on Information Systems*.

Gebauer, J., Shaw, M. & Gribbins, M. (2004). Usage and impact of mobile business: An assessment based on the concepts of task/technology fit. *Proceedings of the 10th America Conference on Information Systems*.

Jarvenpaa, S.L., Lang, K., Reiner, T., Yoko, T., & Virpi, K. (2003). Mobile commerce at crossroads. *Communication of the ACM, 12*(46), 41-44.

Kim, D., & Steinfield, C. (2004). Consumers mobile Internet service satisfaction and their continuance intentions. *Proceedings of the 10th America Conference on Information Systems*.

Kohli, R., & Sherer, S. (2002). Measuring payoff of information technology investments: Research issues and guidelines. *Communications of the Association for Information Systems, 9*(27), 241-268.

Kulatilaka, N., & Venkatraman, N. (2001). Strategic options in the digital era. *Business Strategy Review, 4*(12), 7-15.

Leung, F., & Cheung, C. (2004). Consumer attitude toward mobile advertising. *Proceedings of the 10th America Conference on Information Systems*.

Shim, J., Varshney, U., Dekleva, S., & Knoerzer, G. (2003). Mobile wireless technology and services: Evolution and outlook. *Proceedings of the 9th America Conference on Information Systems*.

Smith, H., Kulatilaka, N., & Venkatramen, N. (2002). New developments in practice III: Riding the wave: Extracting value from mobile technology. *Communications of the Association for Information Systems, 8*(32), 467-481.

Tallon, P., Kauffman, R., Lucas, H., Whinston A., & Zhu, K. (2002). Using real options analysis for evaluating uncertainty investments in information technology: Insights from the ICIS 2001 debate. *Communications of the Association for Information Systems, 9*(27), 136-167.

Tarasewich, P., Nickerson, R.C., & Warkentin, M. (2002). Issues in mobile e-commerce. *Communications of the Association for Information Systems, 8*(3), 41-64.

Tuten, P. (2003). Evaluating information technology investments in an organizational context. *Proceedings of the 9th America Conference on Information Systems*.

Walters, C., & Giles, T. (2000). Using real options in strategic decision making. A *Web Magazine of the Tuck School of Business.* Retrieved from *http://mba.tuck.dartmouth.edu/paradigm/spring2000/*

Xin, X. (2004). A model of 3G adoption. *Proceedings of the 10th America conference on Information Systems*.

Yang, S., Chatterjee, S., & Chan, C. (2004). Wireless communications: Myths and reality. *Communications of the Association for Information Systems, 13*(39).

KEY TERMS

Access Point Device: The device that bridges wireless networking components and a wired network. It forwards traffic from the wired side to the wireless side and from the wireless side to the wired side, as needed.

Investment: An item of value purchased for income or capital appreciation.

M-Commerce: The use of mobile devices to improve performance, create value, and enable efficient transactions among businesses, customers, and employees.

Network Interface Card (NIC): The device that enables a workstation to connect to the network and communicate with other computers. NICs are manufactured by several different companies and come with a variety of specifications that are tailored to the workstation and network's requirements.

NPV: The present value of an investment's future net cash flows minus the initial investment. If positive, the investment should be made (unless an even better investment exists); otherwise, it should not.

Option: By definition, gives the holder the right, but not the obligation, to take ownership of an underlying asset at a future point in time.

Standards: Documented agreements containing technical specifications or other precise criteria that are used as guidelines to ensure that materials, products, processes, and services suit their intended purpose.

UBS (Universal Serial Bus) Port: A standard external bus that can be used to connect multiple types of peripherals (including modems, mice, and network adapters) to a computer.

User Interface: An aspect of a wireless device or a piece of software that can be seen, heard, or otherwise perceived by the human user, and the commands and mechanisms the user uses to control its operation and input data.

Voice Recognition: A technology that enables computers to recognize the human voice, translate it into program code, and act upon the voiced commands.

IT Management Practices in Small Firms

Paul B. Cragg
University of Canterbury, New Zealand

Theekshana Suraweera
University of Canterbury, New Zealand

INTRODUCTION

Computer based information systems have grown in importance to small firms and are now being used increasingly to help them compete. For example, many small firms have turned to the World Wide Web to support their endeavours. Although the technology that is being used is relatively well understood, its effective management is not so well understood. A good understanding is important as the management of IT is an attribute that has the potential to deliver a sustainable competitive advantage to a firm (Mata, Fuerst, & Barney, 1995). This chapter shows that there is no one accepted view of the term "IT management" for either large or small firms. However, the term "management" is often considered to include the four functions of planning, organising, leading, and controlling. This framework can be applied to small firms and specifically to their IT management practices.

BACKGROUND

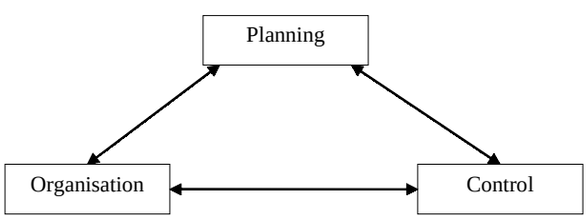
What is meant by the term "IT management"? There are a number of frameworks that can help us understand the concept of IT management. However, most frameworks are based on large firms, with only two specific to small firms, presented in studies by Raymond and Pare (1992) and Pollard and Hayne (1998).

There are three interrelated terms that are frequently used in the literature with respect to the management of computer-based technology: IT management, IS management, and information management.

Two of the terms, Information technology (IT) management and Information systems (IS) management usually refer to the same phenomenon. These terms typically refer to managerial efforts associated with planning, organising, controlling, and directing the introduction and use of computer based systems within an organisation (Boynton et al., 1994). This characterisation is in agreement with the definition of "management" described in classical management literature expressed as a process of four functions, namely planning, organising, leading, and controlling[1] (Schermerhorn, 2004). We see little advantage in attempting to distinguish between IT and IS. Thus, IT management and IS management refer to the same activities, that is, to the organisation's practices associated with planning, organising, controlling, and directing the introduction and use of IT within the organisation.

Table 1 provides examples of the concept of IT management, but before that we should clarify the term information management. It is a term which has frequently been used by authors to refer to two different but related activities. Some conceptualise information management as a process comprised of planning, organisation and control of information resources (see Figure 1 based on Earl, 1989). Thus Earl's information management is the same as IT management, as described above. However, other authors use the term information management to

Figure 1. Earl's model of information management (Earl, 1989)



Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. Key aspects of IT management

Key Issues in IS Management in Small Firms Pollard & Hayne (1998)	Core IS Capabilities Feeny & Willcocks (1998)	IT Best Practices Cragg (2002)	IT Management Processes Luftman (2004)
IS for competitive advantage	IS/IT Leadership	Managers view IT as strategic	Strategic planning and control
IS project management	Business systems thinking	Managers are enthusiastic about IT	Management planning
Software development	Relationship building	Managers explore new uses for IT	Development planning
Responsive IT infrastructure	Architecture planning	New IT systems are customised	Resource planning
Aligning IS Technological change	Making technology work	Firm employs an IT specialist	Service planning
Communication networks	Informed buying	Staff have the skills to customise IS	Project management
Business process redesign	Contract facilitation		Resource control
Educating users	Contract monitoring		Service control
IS human resource	Vendor development		Development and maintenance
			Administration services
			Information services

recognise that organisations have information that needs to be managed. For example, Osterle, Brenner, and Hilbers (1991) claim the fundamental responsibility of information management is to ensure that the enterprise recognizes and harnesses the potential of information as a resource. This view of information management is an important subset of IT management. "IT management" as a broader term, recognises that an organisation has to manage information, as well as hardware, software, people, and processes.

The above discussion defined IT management as practices associated with planning, organising, controlling, and directing the introduction and use of IT within an organisation. Table 1 provides some examples of these practices, based on the work of Cragg (2002), Feeny and Willcocks (1998), Luftman (2004), and Pollard and Hayne (1998),.

Notes for Table 1:

a. Pollard and Hayne (1998) examined the key issues of IT management in small firms in Canada using the Delphi technique. The 10 most critical issues that small firms expect to face in the 1995-2000 era are given above.
b. Feeny and Willcocks (1998) presented nine core IT capabilities based on the experience of large US-based companies. They stated that these capabilities "are required both to underpin the pursuit of high-value-added applications of IT and to capitalise on the external market's ability to deliver cost effective IT services."
c. Cragg (2002) identified six IT management practices that differentiated IT leaders from IT laggards amongst 30 small engineering firms.
d. Luftman (2004) argues that there are 38 IT processes that have to be managed, whatever the size and type of the organisation. Some of these are at a strategic level (long term), some at a tactical level (short term), and others operational (day to day).

MAIN FOCUS OF THE ARTICLE

Two of the sources in Table 1 are based on studies of small firms, such as Cragg (2002) and Pollard and Hayne (1998). These studies show that many IT management processes are similar for both large and small firms, but typically small firms have to manage IT with low levels of internal IT expertise. Thus, small firms often rely heavily on external expertise, as highlighted by several researchers (Fink, 1998; Gable, 1996; Thong, Yap, & Raman 1996). For example, many small firms have no person with a formal IT education. Thong et al. (1996) observed that small businesses rely on consultants and vendors in IT project implementation, and IT effectiveness is positively related to the consultant's effectiveness in such firms.

Numerous studies of IT in small firms have shown that managers within the firm play a key role in both the introduction of new systems and its subsequent success. For example, Caldeira and Ward (2003) concluded that "top management perspectives and attitudes" were one of the two key determinants of IT success in small firms. However, most small firms do not have an IT manager, that is, a person who has IT as their prime managerial responsibility. As a result, many studies have recognised that IT management practices are weak in many small firms, relative to large firms. Fink (1998) argues that the management effort towards IT in small firms is negligible in comparison with that in large firms.

Although the IT managerial processes may differ in small firms, it is not proper to infer that small businesses have absolutely no practices in place for managing their IT. For example, Cragg (2002) pro-

vided many examples of IT management practices in small firms, including coverage of strategic planning, operational planning, and implementation.

The studies of IT management in both large and small firms support the notion that IT management comprises a number of sub-functions. Thus, the general management sub-themes of planning, organising, controlling, and leadership do provide a sound basis for characterising the concept of IT management, in broad terms. However, the differences between IT in small firms and IT in large firms discussed above suggest that the indicators used to characterise planning, organising, and so on, in large firms may not be appropriate in the small business context. Thus it would be unwise to use an instrument tested in large firms, (for instance, for IT planning) in a small firm study before it has been validated on small firms. Our research has provided numerous examples of some important IT management practices in small firms. These are provided in Table 2, grouped under the four sub-dimensions of planning, leadership, controlling, and organising.

Table 2. IT management practices in small firms

Function	Examples of IT Management Practices in Small Firms
IT Planning	Recognising IT planning is an important part of the overall business planning process.
	Maintaining detailed IT plans.
	Using an IT planning process within the firm.
	Designing IT systems to be closely aligned with the overall objectives of the firm.
	Frequent review of IT plans to accommodate the changing needs of the firm.
	Continuous search for and evaluating new IT developments for their potential use in the firm.
	Use of IT systems to improve the firm's competitive position.
IT Leadership	Managers create a vision among the staff for achieving IT objectives
	Managers inspire staff commitment towards achieving IT objectives
	Managers direct the efforts of staff towards achieving IT objectives
	Commitment of the top management to providing staff with appropriate IT training.
	Top management believing that IT is critical to the success of the business.
IT Controlling	Closely monitoring the progress of IT projects.
	Monitoring the performance of IT system(s).
	Having comprehensive procedures in place for controlling the use of IT resources. (e.g., who can use specific software or access specific databases)
	Having comprehensive procedures in place for maintaining the security of information stored in computers.
	Having clearly defined roles and responsibilities for IT development and maintenance in the firm.
	Having formal procedures for the acquisition and/or development of new IT systems
IT Organising	Having staff members devoted to managing the firm's IT resources.
	Having established criteria for selecting IT vendors and external consultants
	Staff participating in making major IT decisions.
	Having a flexible approach to organising IT operations and maintenance.
	Having established criteria for selecting suitable software when acquiring new software.

FUTURE TRENDS

The recent studies of IT in small firms by Caldeira and Ward (2003) and Cragg (2002) show that IT management practices do have a significant influence on IT success in small firms. These studies also show that IT management practices are maturing in many small firms and, in some firms, such practices have become very sophisticated. However, as yet, we have no good way of measuring the maturity or sophistication of management practices in small firms. Present attempts are in the early stages of development as researchers adapt ideas based on instruments used in large firms. For example, Cragg, King and Hussin (2002) focused IT strategic alignment, and Levy and Powell (2000) focused on information systems strategy processes. These instruments need further testing and adaptation.

We also need to better understand the influences on IT management maturity. For example, why have some small firms developed more mature approaches to IT management? What are the factors that have influenced such developments? These lines of enquiry may help us better understand IT cultures within small firms. A better understanding could then unlock ways that could help more small firms use more sophisticated IT: a problem identified by Brown and Lockett (2004).

CONCLUSION

Although there is no one accepted view of the term "IT management" for either large or small firms, the literature indicates that "management" consists of the four functions of planning, organising, leading, and controlling. This framework can be applied to small firms and specifically to their IT management practices. This chapter has provided numerous examples of such IT management practices, based on research in small firms. However, there have been relatively few studies of IT management practices in small firms. This conclusion identifies a significant research opportunity, especially as some believe that IT management has a significant influence on IT success, and can be a source of competitive advantage to small firms.

REFERENCES

Boynton, A.C., Zmud, R.W. & Jacobs, G.C. (1994). The influence of IT management practice on IT use in large organisations. *MIS Quarterly, 18*(3), 299-318.

Brown, D.H. & Lockett, N. (2004). Potential of critical e-applications for engaging SMEs in e-business: A provider perspective. *European Journal of Information Systems,* (4), 21-34.

Caldeira, M.M. & Ward, J.M. (2003). Using resource-based theory to interpret the successful adoption and use of information systems and technology in manufacturing small and medium-sized enterprises. *European Journal of Information Systems, 12*, 127-141.

Cragg, P., King, M. & Hussin, H. (2002). IT alignment and firm performance in small manufacturing firms. *Journal of Strategic Information Systems. 11*, 109-132.

Cragg, P.B. (2002). Benchmarking information technology practices in small firms. *European Journal of Information Systems,* (4), 267-282.

Earl, M.J. (1989). *Management strategies for information technology.* UK: Prentice Hall.

Feeny, D.F. & Willcocks, L.P. (1998). Core IS capabilities for exploring information technology. *Sloan Management Review,* (3), 9-21.

Fink, D. (1998). Guidelines for the successful adoption of information technology in small and medium enterprises. *International Journal of Information Management,* (4), 243-253.

Gable, G.G. (1996). Outsourcing of IT advice: A success prediction model. *1996-Information Systems Conference of New Zealand*, Palmerston North, New Zealand, IEEE.

Levy, M. & Powell, P. (2000). Information systems strategies for small and medium-sized enterprises: An organisational perspective. *Journal of Strategic Information Systems,* (1), 63-84.

Luftman, J.N. (2004). *Managing the information technology resource: Leadership in the information age.* NJ: Pearson Prentice Hall.

Mata, F. J., Fuerst, W.L. & Barney, J.B. (1995). Information technology and sustained competitive advantage: A resource based analysis. *MIS Quarterly,* (5), 487-505.

Osterle, H., Brenner, W. & Hilbers, K. (1993). Total Information Systems Management (Undernehmensfuhrung und Informationssystem:der Ansatz des St. Galler Informationssystem-Managements). R. Boland & R. Hirschheim (Eds.), *Wiley series in information systems*. UK: John Wiley & Sons.

Pollard, C.E. & Hayne, S.C. (1998). The changing face of information system issues in small firms. *International Small Business Journal,* (3), 71-87.

Schermerhorn, J.R. (2004). *Management: An Asia-Pacific perspective*. Milton: John Wiley & Sons.

Thong, J.Y.L., Yap, C. & Raman, K.S. (1996). Top management support, external expertise and information systems implementation in small business. *Information Systems Research,* (2), 248-267.

KEY TERMS

Controlling: Monitoring performance, comparing results to goals, and taking corrective action. Controlling is a process of gathering and interpreting performance feedback as a basis for constructive action and change.

External Support: Assistance from persons outside the firm. Some firms pay for such support by employing a consultant. Other common forms of external support include IS vendors, and advice from peers, that is, managers in other firms.

IT Alignment: How well a firm's information systems are linked to the needs of the business. One way of measuring alignment is to examine how well a firm's business strategy is linked to their IS strategy.

Leading: Guiding the work efforts of other people in directions appropriate to action plans. Leading involves building commitment and encouraging work efforts that support goal attainment.

Management Support: Managers can provide degrees of support for IT. For example, some man-

agers take the lead role as they are keen to see the organisation adopt a new system, for example, the Internet. Other managers may take less active role, for example, by giving approval for financial expenditure but not getting involved in the project.

Organising: Allocating and arranging human and material resources in appropriate combinations to implement plans. Organising turns plans into action potential by defining tasks, assigning personnel, and supporting them with resources.

Planning: Determining what is to be achieved, setting goals, and identifying appropriate action steps. Planning centres on determining goals and the process to achieve them.

Small Firm: There is no universal definition for either of these two terms. Most definitions are based on the number of employees, but some definitions include sales revenue. For example, 20 employees is the official definition in New Zealand, while in North America, a firm with 500 could be defined as a small firm. Another important aspect of any definition of "small firm" is the firm's independence, that is, a small firm is typically considered to be independent, that is, not a subsidiary of another firm.

ENDNOTE

[1] (a) Planning: determining what is to be achieved, setting goals, and identifying appropriate action steps; (b) Organising: allocating and arranging human and material resources in appropriate combinations to implement plans; (c) Leading: guiding the work efforts of other people in directions appropriate to action plans; (d) Controlling: monitoring performance, comparing results to goals, and taking corrective action (Schermerhorn, 2004).

iTV Guidelines

Alcina Prata
Higher School of Management Sciences, Portugal

iTV DEFINITION AND SERVICES

Technology advances ceaselessly, often in the direction of improving existing equipment. Television, for example, has benefited greatly from the emergence and/or transformations that have occurred in a variety of devices, communication platforms, and ways and methods of transmission.

The appearance of computers that store data digitally, the growth of the Internet, which is accessible anytime, anyplace, to anybody (Rosenberg, 2001) and, finally, the appearance of transmission methods that allow for communication in two directions, have led to a new paradigm: interactive television (iTV). iTV, which is a result of the combination of digital television and Internet technology (Nielsen, 1997) in order to deliver a mix of programming, with restricted or open Web access (Chandrashekar, 2001), allows the viewer to interact with an application that is delivered via a digital network simultaneously with the traditional TV signal (Perera, 2002). This means that a concurrent transmission occurs: Namely, the standard program or traditional television broadcast occurs along with the application with interactive elements (Bernardo, 2002). In order to decode the digital information, that is, the abovementioned applications, with interactive elements, a digital adapter must be used by the viewer. This is the so-called set-top box. In order to allow the viewer to interact with the application, a return channel is also needed. The return channel allows the viewer feedback to reach the TV operator.

Several types of services are possible through iTV's principally interactive programs, which offer the possibility of interacting electronically with a normal TV program while it is being broadcast. Other services also include the following:

- Enhanced TV services such as EPGs (electronic programming guides)
- Special services through TV that are made available via the so-called TV sites, namely, weather services, TV shopping, TV banking, games, educational services (t-learning, which is learning through interactive digital TV; Port, 2004), and interactive games amongst others
- Internet browsing and the use of e-mail (Bernardo, 2002; Chambel, 2003)

Different services imply different types and levels of interactivity, which means that iTV may be defined in a multitude of ways (Gill & Perera, 2003). However, what is important to underline is that television and interactivity are "coming together fast" (Bennett, 2004) and, as a nascent phenomenon, iTV is trying to "find its feet, lacking compatibility, interoperability and solid guidelines" (Gill & Perera, 2003, pp. 83-89).

In terms of research areas, the establishment of "solid guidelines" is probably one of the most urgent priorities since so far the largest investments have been in the technological area. For example, a very expensive and time-consuming system developed by Sportvision Inc., USA, (http://www.sportvision.com) for hockey games worked fine in technological terms but "manages to offend even hockey fans with its lack of subtlety" (Television, 2003, pp. 32-35). For Jana Bennett, director of BBC Television, one of the most successful digital iTV operators in Europe that had more than 7.2 million users by the end of 2003 (Quico, 2004), the "biggest challenge ahead will be creative rather than technical. What's needed now is a creative revolution every bit as ambitious as the technical one we have seen" (Bennett, 2004).

Several researchers argue the need of new and personalized services embodying good design (Bennett, 2004; Chorianopoulos, 2003; Damásio, Quico, & Ferreira, 2004; Eronen, 2003; Gill & Perera, 2003; Port, 2004; Prata & Lopes, 2004; Quico, 2004), usability (Gill & Perera, 2003), and subtlety (Television, 2003). These characteristics will be impossible to achieve without specific iTV guidelines based on scientific principles. Thus, we conclude that the next important steps to be taken can be

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

summarized in one sentence: Research must lead to solid guidelines that can be applied in developing creative new personalized services to meet viewers' needs.

iTV GUIDELINES: FINDING THE WAY

In order to produce good iTV interfaces, namely, TV sites and interactive program applications, some specific guidelines need to be followed. However, iTV interface design is still in an embryonic phase (Bernardo, 2003) and, as it is a very recent phenomenon, no specific iTV guidelines have been defined and accepted worldwide. Since iTV uses Internet technology, designers decided to start by focusing their attention on the accepted worldwide Web-site guidelines. However, as the output devices to be used are completely different (the PC [personal computer] versus TV), these Web-site guidelines need to be greatly modified before being applied. Unfortunately, a considerable number of TV sites have already been designed by Web (or ex-Web) designers who were not capable of adapting the above-mentioned guidelines. The result has been poor and inadequate interfaces (Bernardo).

iTV GUIDELINES: TV VS. PC

As previously mentioned, the best starting point for researching new iTV guidelines may be to focus on Web-site guidelines and, after comparing the specific output devices that are going to be used, adapting them. The comparison of the two devices (TV and PC) is a complex and time-consuming process. Since the process does not fall within the scope of this work, only the main aspects are presented.

A brief TV-PC comparison in technical terms allows us to note the following:

- When referring to the TV set, we use the word viewer, and when referring to the PC, we use the word user (Prata, Guimarães, & Kommers, 2004).
- TV implies a broadcast transmission while the PC implies a one-to-one transmission (Bernardo, 2003).
- The TV screen is very different from a traditional PC screen principally in that it has a lower resolution (Bernardo, 2003).
- TV interaction is assured via a TV remote control instead of a mouse, which means that the interface needs to be dramatically adapted for this new and very limited navigational device. The use of remote control implies sequential navigation, whereas interaction with the PC is by means of a mouse, which is much more flexible than a remote control (Bernardo, 2003).
- With TV, all viewers have the same return channel. Thus, content produced for a specific bandwidth will be compatible with the entire audience. Amongst PC users, there are different connection speeds: analogical lines, ISDN (Integrated Services Digital Network), and high-debit connections. This means that each user may achieve a different result (Bernardo, 2003).
- The TV screen has a fixed resolution that viewers are unable to change and that depends on the TV system being used (The PAL [Phase Alternation Line] system used in Portugal, for example, has a resolution of 672x504, which means that the content area must be 640x472 pixels.) The PC screen has a variable resolution that the user is able to change: 1024x768, 800x600, 480x640, and so forth (Bernardo, 2003).
- The TV set easily allows the use of video (the most powerful communication medium) while the PC is still far from handling it easily (Bernardo, 2003; Prata et al., 2004).
- Watching TV is a social activity, and thus, since it is a group phenomenon, it is associated with group interaction (Bernardo, 2003; Masthoff & Luckin, 2002). The PC typically implies individual interaction (Bernardo, 2003).
- On TV, sound and images are of high quality and in real time while, through a PC, sound and images are of lower quality and take some time to arrive (This is an environment where the download time is a variable to be considered.) (Bernardo, 2003)
- Horizontal scrolling is not possible with a TV. At the PC, although not recommended, horizontal scrolling is allowed (Hartley, 1999; Lynch & Horton, 1999; Nielsen, 2000).

- With TV, the opening of new windows in the browser is not possible (everything happens in the same window), while on the PC several different windows may be opened at the same time (Bernardo, 2003).
- With TV, rates charged for advertising time are calculated according to audience ratings, whereas with the PC advertisers pay in accordance with the number of hits, clicks, or page views (Bernardo, 2003).

A brief comparison of TV with the PC in terms of the characteristics of their target populations enables us to see the following:

- TV has a very heterogeneous public (basically everybody, meaning people of all ages and with all types of experience and previous knowledge) while the PC has a more specific and homogeneous public (Bernardo, 2003).
- Almost everyone has a TV set (In Europe, the penetration rate of TV is about 95% to 99%). Far fewer people have PCs and Internet connection than TVs (in Europe the penetration rate of the Internet is approximately 40% to 60% of the population; Bates, 2003).
- With TV, viewers do not work with scroll bars, while with the PC, users are very used to working with them (Bernardo, 2003).

A brief comparison of TV and the PC in terms of viewers' and users' states of mind enables us to see the following:

- TV may provide a pleasant feeling of companionship, while the PC does not necessarily provide this feeling (unless we consider the use of chats; Bernardo, 2003; Prata et al., 2004).
- TV is considered to be a safe environment since it involves no viruses, hackers, crackers, or loss of privacy. Nobody gets into our TV set to steal our information as may happen with Internet-connected computers. The PC with an Internet connection is typically a "hostile" environment with multiple dangers such as those just mentioned (Bernardo, 2003; Prata et al., 2004).
- In general terms, viewers feel more protected while using TV than the Internet because there are special entities responsible for regulating program broadcasts (Bernardo, 2003).
- While watching TV, the viewer is typically seated on a sofa somewhere around three to five meters away from the screen and is usually comfortable and relaxed. While using a PC, the user is very close to the screen, usually seated with the back straight and in a very tense position since he or she needs to handle the mouse (Bernardo, 2003).
- While watching TV, the viewer is less attentive since he is typically in an entertainment environment. While using a PC, the user is more attentive since he or she is typically in a work environment (writing, searching for, or reading information) or in a very interactive entertainment environment playing games (Bernardo, 2003).
- While watching TV, viewers are not expecting to encounter mistakes or problems. However, while using PCs, users are, in general terms, prepared to deal with frequent mistakes or problems (for instance, the need to restart a computer, the time involved with downloads, receipt of illegal operation messages, and so on; Bernardo, 2003).
- While watching TV, viewers are not in a state of "resistance" since viewing is familiar to everyone. While using a PC, users are in a state of greater resistance since not everybody uses the PC and there is a tendency to believe that using it will be difficult (Bernardo, 2003).
- While watching TV, if the interactivity is not instantaneous, viewers will become impatient. The problem is that they are used to changing channels in less than two seconds (more than two seconds will probably cause the viewer to lose concentration; Bernardo, 2003). When using a PC, on the other hand, if the interactivity and content access are not immediate, users are accustomed to waiting. In this case users are prepared to wait for a page or image download for up to 10 seconds (Lynch & Horton, 1999).

iTV GUIDELINES: SPECIFIC PROPOSALS

The study and comparison of the characteristics of the two devices (TV and the PC with Internet connection) briefly presented in the previous section, along with the adaptation of Web-site design guidelines to this new environment (iTV), has led to a wide range of specific iTV guidelines as proposed by several authors. Some of the most important guidelines to be considered when planning, developing, and evaluating iTV interfaces are presented below, separated by categories.

Text Guidelines

- The text pitch used must be 18 minimum in order to be visible from three to five meters away, which is the distance between the viewer and the TV set. Usually, the recommended pitch is 20 for general text and 18 for the observation section(s) or subsection(s). As to the font style, Arial, Helvetica, and Verdana are recommended. Other font styles may be used, but only if embedded as images. However, this solution needs to be carefully considered since the result will be a much heavier file (Bernardo, 2003).
- Small-pitch text embedded in images should be avoided since the browser frequently resizes these images automatically (Bernardo, 2003).
- The text paragraphs must be short in order to not occupy several screens and thus impose the use of scrolling, which is a feature that is hard to handle in iTV (Bernardo, 2003).

Graphics and Background Guidelines

- Rigorous graphics should be avoided since there is always a little toning down (Thin lines may result in some scintillation; Bernardo, 2003).
- Animated graphics, that is to say, graphics with lots of movements, should be avoided (Bernardo, 2003).
- The usage of image maps should be avoided since they are complex to handle on a TV set (Bernardo, 2003).
- The use of very small frames must be avoided since this may result in many differences in the Web page as seen through the PC browser and when seen through the set-top-box browser (Bernardo, 2003).
- It is preferable to use normal graphic buttons with simple words than very graphical buttons full of colours (Bernardo, 2003).
- The TV object (video file embedded in the TV site) should be as large as possible, but the equilibrium between that object and the remaining information (normally textual information) must obviously be kept (Bernardo, 2003).
- When designing a TV site, it is necessary to take into consideration a status bar with a height of 40 pixels. A margin of 16 pixels is recommended for the perimeter of the screen (Bernardo, 2003).
- The background, instead of being an image, should be developed directly in the programming code in order to have less weight. However, if an image needs to be used, it should be simple so that it may be replicated all around the screen without becoming too heavy. Watermarks may also be used since the image only contains one colour (Bernardo, 2003).
- Dark colours should be used as backgrounds. Highly saturated colours such as white should not be used (Bernardo, 2003).

Interactivity Guidelines

- Interactivity may be available in two options: The TV object may be integrated in the Web page, or the contents may be displayed over the television signal (Bernardo, 2003).
- It is essential to bear in mind that the program broadcast is of greatest importance. The rest is secondary and is used to improve the viewer's television experience (Bernardo, 2003).
- The interactive content is supposed to improve the program broadcast without disturbing the viewer's entertainment experience (Bernardo, 2003; Prata et al., 2004).
- The service must be pleasing to the viewer; otherwise, he or she will change the channel (Bernardo, 2003).
- The interface must be easy to understand and allow for easy interaction. A bad design typically forces the viewer to click a large number

of times in order to reach important information. It is important to keep in mind that a large number of clicks does not necessarily mean a very interactive service. Similarly, ease of interaction does not mean less interaction (Bernardo, 2003).

Technical Guidelines

The following information is descriptive only of the Microsoft TV Platform, which is well known worldwide.

- This platform supports the use of the following programming languages: HTML 4.0 (hypertext markup language; full and with some extensions), cascading style sheets (CSS; a subgroup of CSS1 and the absolute positioning of CSS2 [CSS-P], Microsoft TV Jscript, Active X components, and DHTML (Dynamic HyperText Markup Language).
- The platform also permits the integration of Flash 4.0 (or a lower version) animations but with some drawbacks since these animations are very heavy (Bernardo, 2003).
- The platform supports the use of the following file formats: sound (AIFF, WAV, AV, ASF, MP3, and others), image (GIF, JPEG, PNG), video (ASF, ASX), and animation (Flash 4.0 or inferior file formats; Bernardo, 2003).

Other Guidelines

- The dimensions of the TV object must maintain the format 4:3 in order to not distort the television image (Bernardo, 2003).
- Each screen should not take more than three to five seconds to download. However, the ideal time is around two seconds, which is the time it normally takes to change the TV channel (Bernardo, 2003).
- The final design of each screen should occupy a maximum of 100 Kb (Bernardo, 2003).
- Vertical scroll, although possible, should be avoided since it is not practical to navigate via a remote control (however, vertical scroll is used in almost every Web site) (Bernardo, 2003).
- It is important to remember that not all viewers are experienced in the use of Internet scrolling and navigation (Bernardo, 2003).
- The best way of testing a TV site is to use a test population consisting of housewives and/or grandmothers. The critical point is that the usual consumer has to be able to interact with the service using only a remote control. Since such viewers are in the majority, it is essential to capture this specific market of viewers, which consists of people who have probably never used a PC and/or an Internet connection (Bernardo, 2003).
- There is a significant difference between the way we capture the iTV viewer's attention and the way we capture the Internet user's attention. The iTV viewer is used to being entertained, so the challenge will have to be very high in order to capture his or her attention. The quality of the service will also have to be high in order to keep his or her attention (Bernardo, 2003; Masthoff, 2002).

CONCLUSION

According to recent studies, iTV is here to stay. However, since it is a recent phenomenon, additional research is needed, especially with regard to innovative and more personalized services that will be more adapted to viewers' needs. In order to design and develop these new services correctly, new guidelines specifically designed for iTV are needed. The author of the present article has conducted a detailed research study of what should be some of the most important and critical ones and has presented her findings here. However, it will be critical to the success of iTV services in the future that guidelines be continuously developed.

REFERENCES

Bates, P. (2003). *T-learning: Final report*. Report prepared for the European Community. Retrieved on January 3, 2005, from *http://www.pjb.co.uk/t-learning/contents.htm*

Bennett, J. (2004). Red button revolution: Power to the people. *Proceedings of the MIPTV and MILIA 2004*, Cannes, France.

Bernardo, N. (2003). *O guia prático da produção de televisão interactiva*. Centro Atlântico, Lda, Portugal.

Chambel, T. (2003). *Video based hypermedia spaces for learning contexts*. PhD Thesis presented at Lisbon University FCUL, Lisbon, Portugal.

Chandrashekar, A. (2001). *Interactive TV: An approach paper* (White paper). Wipo Technologies.

Chorianopoulos, K. (2003). The virtual channel model for personalized television. *Proceedings of the EuroiTV2003,* Brighton, United Kingdom. Retrieved on January 3, 2005, from *http://www.brighton.ac.uk/interactive/euroitv/euroitv03/Papers/Paper7.pdf*

Damásio, M., Quico, C., & Ferreira, A. (2004, March). Interactive television usage and applications: The Portuguese case-study. *Computer & Graphics Review.*

Eronen, L. (2003). User centered research for interactive television. *Proceedings of the EuroiTV2003,* Brighton, United Kingdom. Retrieved on January 3, 2005, from *http://www.brighton.ac.uk/interactive/euroitv/euroitv03/Papers/Paper1.pdf*

Gill, J., & Perera, S. (2003). Accessible universal design of interactive digital television. *Proceedings of the EuroiTV2003,* Brighton, United Kingdom. Retrieved on January 3, 2005, from *http://www.brighton.ac.uk/interactive/euroitv/euroitv03/Papers/Paper10.pdf*

Hartley, K. (1999). Media overload in instructional Web pages and the impact on learning. *Educational Media International, 36*, 45-150.

Lynch, P., & Horton, S. (1999). *Web style guide: Basic design principles for creating Web sites*. CT: Yale University Press.

Masthoff, J. (2002). Modeling a group of television viewers. *Proceedings of the TV'02 Conference*, (pp. 34-42). Retrieved January 3, 2005, from *http://www.it.bton.ac.uk/staff/jfm5/FutureTV02paper.pdf*

Masthoff, J., & Luckin, R. (2002). Workshop future TV: Adaptive instruction in your living room. *Proceedings of the TV02 Conference,* (pp. 1-3). Retrieved on January 3, 2005, from *http://www.it.bton.ac.uk/staff/jfm5/FutureTV*

Nielsen, J. (1997). *TV meets the Web*. Retrieved on January 3, 2005, from *http://www.useit.com/alertbox/9701.html*

Nielsen, J. (2000). *Designing Web usability*. Indianapolis, IN: New Riders Publishing.

Perera, S. (2002). Interactive digital television (Itv): The usability state of play in 2002. *Scientific and technological reports.* Retrieved on January 3, 2005, from *http://www.tiresias.org/itv/itv1.htm*

Port, S. (2004). *Who is the inventor of television*. Retrieved on January 5, 2005, from *http://www.physlink.com/Education/AskExperts/ae408.cfm*

Prata, A., & Lopes, P. (2004). Online multimedia educational application for teaching multimedia contents: An experiment with students in higher education. In P. Darbyshire (Ed.), *Instructional technologies: Cognitive aspects of online programs* (pp. 31-72). Hershey, PA: Idea Group Publishing.

Prata, A., Guimarães, N., & Kommers, P. (2004). e-iTV multimedia system: Generator of online learning environments through interactive television. *Proceedings of the conference INTERACÇÃO 2004 (First National Conference on Human Computer Interaction),* Lisbon, Portugal.

Quico, C. (2004). Televisão digital e interactiva: O desafio de adequar a oferta às necessidades e preferências dos utilizadores. *Proceedings of the Televisão Interactiva: Avanços e Impactos conference,* Lisbon, Portugal.

Rosenberg, M. (2001). *E-learning: Strategies for delivering knowledge in the digital age.* New York: McGraw-Hill.

Television. (2003, November). *IEEE Spectrum*, 32-35.

KEY TERMS

Broadcast: A transmission to multiple unspecified recipients.

Digital Television: The new generation of broadcast television transmissions. These are of better quality than the traditional analogical broadcasts and will presumably replace them.

File Formats: The way a file stores information—the way in which a file is saved. The file format depends on the content that is being stored, the application that is being used, and the compression algorithm that is being used.

Guidelines: Design and development principles that must be followed in order to achieve a good application.

Programming Language: A formal language in which computer programs are written. The definition of a particular language consists of both syntax (which refers to how the various symbols of the language may be combined) and semantics (which refers to the meaning of the language constructs).

Set-Top Box: Device used in order to convert digital information received via interactive television.

TV Object: A video file embedded in a TV site, normally surrounded by other elements such as textual information.

Leadership Competencies for Managing Global Virtual Teams

Diana J. Wong-MingJi
Eastern Michigan University, USA

INTRODUCTION

The demand for leadership competencies to leverage performance from global virtual teams (GVTs) is growing as organizations continue to search for talent, regardless of location. This means that the work of virtual leaders is embedded in the global shifting of work (Tyran, Tyran & Shepherd, 2003). The phenomenon began with the financial industry as trading took place 24/7 with stock exchanges in different time zones. It is expanding into other industries such as software programming, law, engineering, and call centers. GVTs support the globalization of work by providing organizations with innovative, flexible, and rapid access to human capital. Several forces of competition contribute to the increasing adoption of GVTs, including globalizing of competition, growing service industries, flattening of organizational hierarchies, increasing number of strategic alliances, outsourcing, and growing use of teams (Pawar & Sharifi, 1997; Townsend, DeMarie & Hendrickson, 1998). The backbone of GVTs is innovation with computer-mediated communication systems (CMCSs). Advances with CMCSs facilitate and support virtual team environments.

Leaders of GVTs have a pivotal role in mediating between the internal team processes and the external environment. Leadership competencies also are necessary to keep up with the evolving demands placed on GVTs. Previously, GVTs focused primarily on routine tasks such as data entry and word processing. More recently, the work of GVTs began to encompass non-routine tasks with higher levels of ambiguity and complexity. By tackling more strategic organizational tasks such as launching multinational product, managing strategic alliances, and negotiating mergers and acquisitions, GVTs contribute higher added value to a firm's competitive advantage. As a result, leadership competencies for GVTs become more important in order to maximize the performance of GVTs.

Leadership competencies encompass knowledge, skills, abilities, and behaviors. The following discussion reviews the context, roles, and responsibilities of managing GVTs, identifies five broad categories of GVT leadership competencies, and outlines significant future trends.

BACKGROUND

In order to address specific leadership competencies for GVTs, it is important to understand the virtual workplace context. "Global virtual teams being a novel organizational design, it is very important to maximize the fit between team design and their stated intent" (Prasad & Akhilesh, 2002, p. 104). Currently, many organizations are deploying the use of GVTs much more rapidly than the collective understanding of their unique characteristics, dynamics, and processes. Anecdotal evidence exists about the difficulties and poor performance of GVTs. But the expectations of flexibility, accessing expertise regardless of geographical location, and speed of fulfilling organizational goals continue to drive the growth of GVTs (Gibson & Cohen, 2003).

GVTs have similarities and differences when compared with traditional teams (Maznevski & Chudoba, 2000). The similarities include being guided by shared goals, working on interdependent tasks, and sharing responsibilities for outcomes. The differences are the collocation and synchronous communication of traditional teams vs. geographical dispersion and often asynchronous communication for virtual teams. The stability of GVTs depends on the project and the team's role in fulfilling the organizational purpose. Thus, GVT leaders may be working with a project orientation or indefinite per-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

petual organizational responsibilities, which shape the lifecycle of the team.

Effective GVT leaders must manage magnified ambiguities and complexities compared to traditional team leaders. Prasad and Akhilesh (2002) define a GVT as "a team with distributed expertise and that spans across boundaries of time, geography, nationality, and culture" (p. 103). They address a specific organizational goal with enhanced performance and operate with very little face-to-face interaction and predominantly computer mediated and electronic communication. As a result, leaders of GVTs need to address unique challenges that stem from spatial distances, asynchronous communication, multicultural dynamics, and national boundaries in a virtual environment.

Established research findings on teams indicates that leaders have a critical influence on team performance outcomes (Bell & Kozlowski, 2002; Fjermestad & Hiltz, 1998-1999; Kayworth & Leidner, 2001-2002). In general, team leaders have two critical functions: team development and performance management. Some general leadership tasks for managing teams include developer of team processes, facilitators of communications, and final arbiter for task completion (Duarte & Tennant-Snyder, 1999). Bell & Kozlowski (2002) offer a typology of virtual teams based on four characteristics—temporal distribution, boundary spanning, lifecycle, and member roles—that are mediated by task complexity. These characteristics imply that effective management of GVTs requires a portfolio of leadership competencies to address the following responsibilities: (1) provide clear direction, goals, structures, and norms to enable self regulation among team members; (2) anticipate problems; (3) monitor the environment and communicate changes to inform team members; (4) design back-up plans to buffer changes in environmental conditions; (5) develop feedback opportunities into team management structure for regular performance updates; (6) diagnose and develop appropriate team development through a virtual medium; (7) diagnose the translation of self-regulation methods across different boundaries; (8) modify behaviors and actions according to the particular situations to support the communication of worldviews among team members and build a third culture; and (9) identify and communicate team member roles to create role networks.

An important component of the GVT leader's work environment is the virtual "rooms" for the team's interactions. A wide range of products offers differing capabilities. For example, Groove Client 2.5 and Enterprise Management from Groove Networks, Workgroup Suite 3.1 from iCohere, and eRoom 7.0 from Documentum are products that facilitate how virtual teams can navigate through cyberspace (Perey & Berkley, 2003). Large firms in the auto industry use a commercial B2B product called ipTeam from NexPrise to support collaboration among geographically dispersed engineering team members. IBM offers the IBM Lotus Workplace Team Collaboration 2.0. Free Internet downloads such as NetMeeting from Microsoft also are available to facilitate virtual meetings. Competitors include FarSite from DataBeam Corp, Atrium from VocalTec Communications Ltd., ProShare from Intel Corp, and Conference from Netscape. The list of available CMCS products continues to grow and improve with more features that attempt to simulate face-to-face advantages. As a result, part of managing GVTs includes evaluating, selecting, and applying the most appropriate CMCS innovations to support team interactions. Adopting CMCS needs to account for work locations, members involved, technological standardization, work pace, work processes, and nature of work in the organization. In sum, a GVT leadership portfolio must be able to manage CMCSs, diverse team members, team development, and work flow processes.

GVT LEADERSHIP COMPETENCIES

Competencies for GVT leaders can be classified into five broad categories: CMCS proficiency, work process design, cross-cultural competencies, interpersonal communication, and self-management. The five groups of competencies are interrelated. For example, a high degree of expertise with CMCSs without the necessary interpersonal communication competencies likely will lead to conflicts, absentees, and negative productivity.

First, GVT leaders need to have technical proficiency with innovations in CMCS in order to align the most appropriate technological capabilities with organizational needs. Technical knowledge of CMCSs and organizational experience enables GVT leaders

to align technology with strategic organizational goals. Organizational experience provides GVT leaders with insights regarding the organizational work task requirements, strategic direction, and culture. This tacit knowledge is rarely codified and difficult to outsource compared to explicit knowledge. This implies that firms should provide training and professional development for leaders to increase CMCS proficiency.

Second, GVT leaders require work process design competencies to manage the workflows. Managing global virtual workflows depends on leadership skills to structure teams appropriately for subtasks, monitor work progress, establish expectations, maintain accountability, build a cohesive team, motivate team members, create trust, develop team identity, and manage conflicts (Montoya-Weiss, Massey & Song, 2001; Pauleen & Yoong, 2001; Piccoli & Ives, 2003). GVT leaders also need to devote considerable attention to performance management, especially in prototypical teams where there may be information delays and members are decoupled from events. GVT leaders can employ temporal coordination mechanisms to mitigate negative effects of avoidance and compromise in conflict management behavior on performance (Montoya-Weiss, Massey & Song, 2001). During the launching of teams, GVT leaders need to use appropriate team building techniques (e.g., discussion forums) to become acquainted and to establish positive relationships (Ahuja & Galvin, 2003; Prasad & Akhilesh, 2002). The lifecycle of virtual teams tends to proceed through four stages of group development that entails forming with unbridled optimism, storming with reality shock, norming with refocus and recommitment, and performing with a dash to the finish (Furst et al., 2004). The lifecycle of virtual teams influences the development of team spirit and identity, which is more important with continuous virtual team lifecycle. Its membership is relatively more stable compared to temporary projects. Task complexity places constraints on team structure and processes (Prasad & Akhilesh, 2002). Relatively simple tasks have less need for stable internal and external linkages, common procedures, and fixed membership, compared to more complex tasks. Leaders need to assert flexible, collegial authority over tasks and act as empathetic mentors to create collaborative connections between team members (Kayworth & Leidner, 2001). In sum, managing the work process design requires dealing with paradoxes and contradictions to integrate work design and team development.

Third, GVT leaders also require cross-cultural competencies, more specifically identified as global leadership competencies. "Successful virtual team facilitators must be able to manage the whole spectrum of communication strategies as well as human and social processes and perform these tasks across organizational and cultural boundaries via new [information and communication technologies]" (Pauleen & Yoong, 2001, p. 205). Developing global leadership competencies entail a sequence from ignorance, awareness, understanding, appreciation, and acceptance/internalization to transformation (Chin, Gu & Tubbs, 2001). The latter stages involve development of relational competence to become more open, respectful, and self-aware (Clark & Matze, 1999). Understanding cultural differences helps to bridge gaps in miscommunication. Identifying similarities provides a basis for establishing common grounds and interpersonal connections among team members. Leaders who are effective in leading across different cultures have relational competence to build common grounds and trust in relationships (Black & Gregersen, 1999; Gregersen, Morrison & Black, 1998; Manning, 2003). By increasing trust, leaders can connect emotionally with people from different backgrounds to create mutually enhancing relationships (Holton, 2001; Jarvenpaa & Leidner, 1999). The connections are critical to construct a high-performing team (Pauleen, 2003). A key to cross-cultural leadership competencies for GVTs is projecting them into a virtual environment. This is related to CMCS proficiency, which supports the communication cross-cultural competencies in a virtual environment. Cross-cultural competencies also are closely interrelated with both interpersonal communication competencies and self-management to effectively lead GVTs.

Fourth, interpersonal communication competencies do not necessarily encompass cross-cultural competencies. But cross-cultural competencies build upon interpersonal communication competencies. Strong interpersonal communication enables GVT leaders to span multiple boundaries to sustain team relationships (Pauleen, 2003). An important communication practice is balancing the temporal di-

mension and rhythm of work to stay connected (Maznevski & Chudoba, 2000; Saunders, Van Slyke & Vogel, 2004). Interpersonal communication competencies for GVT leaders need to focus on the human dimension. For example, GVT leaders need to be conscious of how they "speak," listen, and behave non-verbally from their receiver's perspective without the advantage of in-the-moment, face-to-face cues. This provides the basis for moving from low to higher levels of communication—cliché conversation, reporting of facts about others, sharing ideas and judgments, exchanging feelings and emotions, and peak communication with absolute openness and honesty (Verderber & Verderber, 2003). Interpersonal communication skills of GVT leaders should, at a minimum, support the exchange of ideas and judgments. When GVT leaders demonstrate "active listening" online, team members likely will move toward higher levels of communication. Active listening in GVTs can be demonstrated with paraphrasing, summarizing, thoughtful wording, avoiding judgment, asking probing questions, inviting informal reports of progress, and conveying positive respectful acknowledgements. Another aspect of interpersonal communication competencies for GVT leaders is establishing netiquette, which establishes ground rules and team culture. GVT leaders can strategically develop their interpersonal communication competencies to socialize team members, build team connections, motivate team commitment, resolve conflicts, and create a productive team culture to achieve high performance outcomes (Ahuja & Galvin, 2003; Kayworth & Leidner, 2001-2002).

Finally, a GVT leader's self-management competencies fundamentally influence the development of the four competencies. GVT leaders need to manage their self-assessment and development to acquire a portfolio of competencies. A high level of emotional intelligence enables GVT leaders to engage in self-directed learning for personal and professional development. Self-management refers to adaptability in dealing with changes, emotional self-control, initiative for action, achievement orientation, trustworthiness, and integrity with consistency among values, emotions and behavior, optimistic view, and social competence (Boyatzis & Van Oosten, 2003). The development of GVT leaders with self-management can positively influence team performance by rectifying areas of their own weaknesses and reinforcing their strengths.

In summary, GVTs provide organizations with an important forum for accomplishing work and gaining a competitive advantage in global business. Technological innovations in CMCSs provide increasingly effective virtual environments for team interactions. A critical issue focuses on the GVT leader with the necessary portfolio of competencies. Research and understanding of leadership competencies for managing GVTs are at a nascent stage of development.

FUTURE TRENDS

Researchers need to delve into this organizational phenomenon to advance best practices for multiple constituents and help resolve existing difficulties with GVTs. Understanding leadership competencies for managing GVTs depends on a tighter coupling in the practice-research-practice cycle. Given turbulent competitive environments and more knowledge-based competition, research practices need to keep up with the rapid pace of change. At least three important trends about GVTs need to be addressed in the future.

First, GVTs will continue to grow in strategic importance. An important implication is that GVTs will face greater complexities and ambiguities. Furthermore, GVT leaders will have little or no contextual experience with their team members' locations. This is a significant shift when globe-trotting managers often have face-to-face time with their team members in different locations. Thus, the need to create authentic emotional connections and accomplish the task at hand through multiple CMCSs will continue to be important

Second, another important trend is the rapid pace of technological innovations in telecommunications. New developments will create more future opportunities. For example, advances with media-rich technologies enable communication that narrows the gap between virtual and face-to-face interactions. However, there is little understanding about the relationship between technological adoption and team members from different cultural backgrounds. Given cultural differences, an important consideration would be how people will relate to technological innova-

tions. This has implications for how leaders will manage GVTs. This research issue also has implications for firms engaged in developing CMCSs, because it will affect market adoption.

Last, although not least, organizations also will need to keep pace with the growth of GVTs by developing supporting policies, compensation schemes, and investments. GVT leaders can make important contributions to facilitate organizational development and change management.

The existing GVT literature has some preliminary theoretical developments that require rigorous empirical research. Future research needs to draw from intercultural management, organization development (OD), and CMCSs with interdisciplinary research teams. OD researchers and practitioners will provide an important contribution to different levels of change—individual, groups and teams, organizational, and interorganizational—as managers and organizations engage in change processes to incorporate GVTs for future strategic tasks.

CONCLUSION

The use of global virtual teams is a relatively new organizational design. GVTs allow organizations to span time, space, and organizational and national boundaries. But many organizational GVT practices have a trial and error approach that entails high costs and falls short of fulfilling expectations. The cost of establishing GVTs and their lackluster performance creates a demand for researchers to figure out how to resolve a range of complex issues. An important starting point is with the leadership for managing GVTs. Developing a balanced portfolio of five major leadership competencies—CMCS proficiency, work process and team designs, cross-cultural competence, interpersonal communication, and self-management—increases the likelihood of achieving high performance by GVTs.

REFERENCES

Ahuja, M.K., & Galvin, J.E. (2003). Socialization in virtual groups. *Journal of Management*, *29*(3), 161-185.

Bell, B.S., & Kozlowski, S.W. (2002). A typology of virtual teams: Implications for effective leadership. *Group and Organization Management*, *27*(1), 14-49.

Black, J.S., & Gregersen, H.B. (1999). The right way to manage expats. *Harvard Business Review*, *77*(2), 52-59.

Boyatzis, R., & Van Oosten, E. (2003). A leadership imperative: Building the emotionally intelligent organization. *Ivey Business Journal*, *67*(2), 1-6.

Bueno, C.M., & Tubbs, S.L. (2004). Identifying global leadership competencies: An exploratory study. *Journal of American Academy of Business*, *5*(1/2), 80-87.

Chin, C., Gu, J., & Tubbs, S. (2001). Developing global leadership competencies. *Journal of Leadership Studies*, *7*(3), 20-31.

Clark, B.D., & Matze, M.G. (1999). A core of global leadership: Relational competence. *Advances in Global Leadership*, *1*, 127-161.

Duarte, N., & Tennant-Snyder, N. (1999). *Mastering virtual teams: Strategies, tools, and techniques that succeed.* San Francisco, CA: Jossey-Bass.

Fjermestad, J., & Hiltz, S.R. (1998-1999). An assessment of group support systems experiment research: Methodology and results. *Journal of Management Information Systems*, *15*(3), 7-149.

Furst, S.A., Reeves, M., Rosen, B., & Blackburn, R.S. (2004). Managing the life cycle of virtual teams. *Academy of Management Executive*, *18*(2), 6-20.

Gibson, C.B., & Cohen, C.B. (Eds.) (2003). *Virtual teams that work: Creating conditions for virtual team effectiveness.* San Francisco, CA: Jossey-Bass.

Gregersen, H.B., Morrison, A.J., & Black, J.S. (1998). Developing leaders for the global frontier. *Sloan Management Review, 40*(1), 21-33.

Holton, J.A. (2001). Building trust and collaboration in a virtual team. *Team Performance Management*, *7*(3/4), 36-47.

Jarvenpaa, S.L., & Leidner, D.E. (1999). Communication and trust in global virtual teams. *Organization Science*, *10*(6), 791-815.

Kayworth, T.R., & Leidner, D.E. (2001-2002). Leadership effectiveness in global virtual teams. *Journal of Management Information Systems*, *18*(3), 7-31.

Manning, T.T. (2003). Leadership across cultures: Attachment style influences. *Journal of Leadership and Organizational Studies*, *9*(3), 20-26.

Maznevski, M.L., & Chudoba, K.M. (2000). Bridging space over time: Global virtual team dynamics and effectiveness. *Organization Science*, *11*(5), 473-492.

Montoya-Weiss, M.M., Massey, A.P., & Song, M. (2001). Getting it together: Temporal coordination and conflict management in global virtual teams. *Academy of Management Journal*, *44*(6), 1251-1262.

Pauleen, D.J. (2003). Leadership in a global virtual team: An action learning approach. *Leadership & Organization Development Journal*, *24*(3), 153-162.

Pauleen, D.J., & Yoong, P. (2001). Relationship building and the use of ICT in boundary-crossing virtual teams: A facilitator's perspective. *Journal of Information Technology*, *16*, 205-220.

Pawar, K.S., & Sharifi, S. (1997). Physical or virtual team collocation: Does it matter? *International Journal of Production Economics*, *52*, 283-290.

Perey, C., & Berkley, T. (2003). Working together in virtual facilities. *Network World*, *20*(3), 35-37.

Piccoli, G., & Ives, B. (2003). Trust and unintended effects of behavior control in virtual teams. *MIS Quarterly*, *27*(3), 365-395.

Prasad, K., & Akhilesh, K.B. (2002). Global virtual teams: What impacts their design and performance? *Team Performance Management*, *8*(5/6), 102-112.

Saunders, C., Van Slyke, C., & Vogel, D. (2004). My time or yours? Managing time visions in global virtual teams. *Academy of Management Executive*, *18*(1), 19-31.

Townsend, A., DeMarie, S., & Hendrickson, A. (1998). Virtual teams: Technology and the workplace of the future. *Academy of Management Executive*, *12*(3), 17-29.

Tyran, K.L., Tyran, C.K., & Shepherd, M. (2003). Exploring emerging leadership in virtual teams. In C.B. Gibson, & C.B. Cohen (Eds.), *Virtual teams that work: Creating conditions for virtual team effectiveness*, (pp. 183-195). San Francisco: Jossey-Bass.

Verderber, R.F., & Verderber, K.S. (2003). *Interact: Using interpersonal communication skills*. Belmont, CA: Wadsworth.

KEY TERMS

Asynchronous Communication: Information exchanges sent and received at different times, often taking place in geographically dispersed locations and time zones.

CMCS: Computer-mediated communication system includes a wide range of telecommunication equipment such as phones, intranets, Internets, e-mail, group support systems, automated workflow, electronic voting, audio/video/data/desktop video conferencing systems, bulletin boards, electronic whiteboards, wireless technologies, and so forth to connect, support, and facilitate work processes among team members.

Colocation: Team members sharing the same physical location, which allows for face-to-face interaction.

Emotional Intelligence: A set of competencies that derive from a neural circuitry emanating in the limbic system. Personal competencies related to outstanding leadership include self-awareness, self-confidence, self-management, adaptability, emotional self-control, initiative, achievement orientation, trustworthiness, and optimism. Social competencies include social awareness, empathy, service orientation, and organizational awareness. Relationship management competencies include inspirational leadership, development of others, change catalyst, con-

flict management, influence, teamwork, and collaboration.

Human Capital: The knowledge, skills, abilities, and experiences of employees that provide value-added contributions for a competitive advantage in organizations.

Netiquette: This is a combination of the words "etiquette" and "Internet" ("net," for short). Netiquette is rules of courtesy expected in virtual communications to support constructive interpersonal relationships in a virtual environment.

Synchronous Communication: Information exchanges taking place in the same space and time, often face-to-face.

Temporal Coordination Mechanism: A process structure imposed to intervene and direct the pattern, timing, and content of communication in a group.

Learning Networks

Albert A. Angehrn
Center for Advanced Learnimg Technologies, INSEAD, France

Michael Gibbert
Bocconi University, Italy

INTRODUCTION

Herb Simon once said that "all learning takes place inside individual human heads[;] an organization learns in only two ways: (a) by the learning of its members, or (b) by ingesting new members who have knowledge the organization didn't previously have" (as cited in Grant, 1996, p. 111). What Simon seems to be implying is that while organizational learning can be seen as linked to the learning of individuals, these individuals need to be employed by the organization intending to appropriate the value of learning.

We partially agree. Take one of the most fundamental processes—learning—and combine it with one of the most powerful processes to create and distribute value—networks. What emerges is the concept of learning networks (LNs). LNs come in many forms. Two generic forms of LNs stand out. First, LNs that focus on learning and knowledge-sharing processes *within* one organization. This perspective is endorsed by Herb Simon and is also at the heart of knowledge management in that it understands learning as the sharing of knowledge among employees of the same company (e.g., Davenport & Prusak, 1998; von Krogh & Roos, 1995). The internal perspective on learning has its roots in theories of organizational learning in that it sees learning as a process that helps the organization maintain a competitive advantage by careful management of employee's knowledge (Senge, 1990).

But a second form of LNs, which focuses on knowledge sharing *between* organizations, comes to mind. This perspective has its roots in the area of interorganizational collaboration. Interfirm collaborations broadly refer to a variety of interorganizational relationships such as joint development agreements, equity joint ventures, licensing agreements, cross-licensing and technology sharing, customer-supplier partnerships, and R&D (research and development) contracts (e.g., Dyer & Singh, 1998). Researchers have two streams of thought. One focuses on vertical collaboration, that is, customer-supplier relationships that are characterized by legally binding contracts (e.g., Dyer & Nobeoka, 2000). While most literature focuses on those interorganizational relationships that are specified in formal agreements, the knowledge exchange may take place in social networks that are governed by shared norms of the exchange instead of legally binding contracts (Liebeskind, Oliver, Zucker, & Brewer, 1996; Powell, 1998; Powell, Koput, & Smith-Doerr, 1996).

It is on this second stream of thought where we put the emphasis in this article. Four objectives are pursued. First, we intend to define the concept of LNs by way of comparing it with related constructs on both the intra-organizational and interorganizational levels. Second, we trace important developments in the competitive environment that seem to lead to an increasing importance of LNs as we interpret them. Third, and most importantly, we outline what we call the three key challenges (cf. Gibbert, Angehrn, & Durand, in press) that seem to characterize LNs. Finally, we outline important future trends that seem to shift the emphasis among the three key challenges. Here, we briefly preview these three key challenges:

- **"Real" vs. virtual forms of interaction:** Individual members of LNs may interact directly (i.e., person to person) and virtually (i.e., through technology-mediated channels). It is unclear, however, which form of collaboration is most efficient in the learning process.
- **Collaboration vs. competition for learning outcomes:** This arises since LNs involve horizontal collaboration, that is, collaboration among competitors, and because there are typically no formal, legally binding contracts to govern the collaboration.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

- **Value creation vs. value appropriation:** A related issue is the extent to which organizations in an LN may be subject to free-riding behavior.

BACKGROUND

The emergence of LNs should be seen against the background of a number of shifts in the institutional, business, and broader societal environments (e.g., Grant, 1996; Spender, 1996a, 1996b; Stewart, 1998). Leibold, Probst, and Gibbert (2002) list a number of major forces causing significant shifts in strategic management thinking and implementation. The main shifts involved in the emergence of LNs are from

- bureaucracies to networks,
- training and development to learning, and
- competitive to collaborative thinking.

Shift from Bureaucracies to Networks

The traditional hierarchical designs that served the industrial era are not flexible enough to harness the full intellectual capability of an organization. Much more unconstrained, fluid, networked organizational forms are needed for effective, modern decision making. The strategic business units (SBUs) of the Alfred P. Sloan era have given way to the creation and effective utilization of strategic business networks (SBNs) by a given enterprise. Progressive organizations establish strategic business systems (SBSs) with multiple networks, interdependent units, and dual communications. The reality is that effective organizations are neither hierarchical nor networked, but a blend of both. Based on a company's traditions and values, different priorities would be placed on the management spectrum. The important thing is that there is flexibility built into the managerial system to capitalize on opportunities while simultaneously ensuring proper responsibility and accountability. This notion of constrained freedom is more complex than it appears, but holds significant creativity and innovation benefits for the enterprise.

Shift from Training and Development to Learning

The role of education has become paramount in all organizations—public and private. However, the change has been from a passive orientation with a focus on the trainer and the curriculum to an active perspective that places the learner at the heart of the activity. In fact, learning must occur in real time in both structured and informal ways. Detailed curriculums have given way to action research by teams as the best way to advance the knowledge base. The new lens requires one to realize the real-time value of learning—in the classroom, on the job, and in all customer and professional interactions. Learning is the integral process for progress. It is an investment rather than a perceived expense to the organization. The knowledge that one creates and applies is more important than the knowledge one accumulates. New techniques, such as collaborative teams and action research, can be easily incorporated into the culture.

Shift from Competitive to Collaborative Thinking

We live in an era dominated by competitive-strategy thinking, one that produces only win-lose scenarios. Even in a cooperative environment, parties divide up the wealth to create a win-win situation. The pie, however, often remains the same. With a collaborative approach, symbiosis creates a larger pie to share or more pies to divide. Alliances of every dimension are the natural order of the day in the realization that go-it-alone strategies are almost always suboptimal. The last decade has been bursting with institutionalized examples of competitive strategy. It is time to remove the barriers to progress and to establish mechanisms of communication, collaboration, and partnership that transcend current practice. The emerging collaborative practices among traditional competitors, for example, supply-chain collaboration between GM, Ford, and Daimler Chrysler in the automotive industry, illustrate this shift to collaborative learning and strategy.

THE THREE KEY CHALLENGES IN LEARNING NETWORKS

The three key challenges outlined in the introduction are at the heart of interorganizational collaboration involving competing firms.

An analysis[1] of the key governance and value-creation processes of LNs has helped us identify three key challenges of learning in networks: real vs virtual forms of interaction, collaboration and competition in the learning process, and value creation and appropriation in networks.

Real vs. Virtual Forms of Interaction

An interesting development is the inclusion of information technology to facilitate learning in networks. Recent evidence of the inclusion of information technology as a facilitator in LNs includes e-learning and communities of practice that are globally dispersed. The key advantage of information technology in these contexts is efficiency, that is, driving down the cost of the communication and distribution of knowledge. However, despite this promise, the usage of information technology as an enabler to learning in networks poses significant challenges in terms of issues such as "direct touch," building trust, capturing the attention of the members of a learning network, and sustaining learning in computer-mediated, distributed environments. Furthermore, IT has enabled the emergence of new forms of distributed, collaborative learning and knowledge creation as witnessed, for example, by the way open-source communities operate. However, the applicability of this model of learning (by doing) in open-source networks seem limited to the software-development realm, and its promise in contexts other than software development is still an open question.

Collaborating vs. Competing in Learning

LNs draw their value from the collaborative attitude of their members. The collaborative attitude seems to be a function of how well members "speak the same language" (i.e., share the same ontology). However, sharing the same ontology usually means that knowledge-sharing partners operate in similar industries, and even in similar stages of the value chain (E.g., engineers from a company in the automotive-supplier industry talk to engineers from another company). In other words, the learning potential is greatest when interacting parties are competitors. This suggests that an appropriate balance be sought between collaboration and competition, dealing with issues of free-riding problems, non-sharing behavior, and especially the unintentional transfer of knowledge while learning in inter-organizational networks.

Distributed Value Creation vs. Focused Value Appropriation

LNs potentially enable the emergence of different types of value, in particular, knowledge exchange, knowledge creation, and synergies, leading to intellectual capital, social capital, and the development of individual competencies of members. But it is nevertheless still unclear to what extent such value-creation sources can be linked to more traditional (and accepted) performance indicators. In other words, it still seems an open question how the value created in the network can be quantified using tangible rather than intangible indicators. Furthermore, the problematic quantification leads to an additional challenge: Which mechanisms have to be put in place to guarantee a fair redistribution of the value created within such a network? (E.g., how can companies cooperating with their customers within the learning network redistribute the value thus created fairly to customers as the co-creators of this value?)

FUTURE TRENDS

Are the three key challenges equally important, or will there be shifts in emphasis over time? Based on our research, we expect a shift in importance toward the nature of interaction as summarized in the first challenge, virtual vs real forms of interaction.

Is High Touch Better Than High Tech?

In the traditional line of thinking, high tech is useful to save money but does not seem fully satisfactory

in all instances, and it must therefore be enhanced by some high-touch elements. High touch in this context means either (a) at some stage(s) in the LN formation, there has to be a moment when members meet in real time and space, or (b) an LN's communication process can be enriched by interactive technologies that simulate high touch (e.g., teleconferencing, etc.).

The underlying assumption is that somehow we, as human beings, prefer to meet in real time and space (e.g., Zuboff, 1988), and that information technology as a vehicle for transmitting knowledge in some way or other deprives us of the richness of shared social experience. Contributing to this assumption is that some forms of knowledge, particularly tacit knowledge (e.g., Polanyi, 1958), seems to require the sustained collaboration between human beings, in other words, direct contact. Such tacit knowledge, which is often intricately bound to the individual experience, is in many ways more art than science, and is typically not expressible in virtual interaction, say, in e-mails. For example, it seems hard to learn how to cook a Thanksgiving turkey from reading cookbooks or joining a Thanksgiving newsgroup. The apprenticeship system in Europe pays witness to this form of learning by doing where a master craftsman passes on his or her art to the apprentice. Most of the arts use this approach to learning.

Or Is High Tech Better Than High Touch?

What if we let go of this preconception? What if we think that high touch is more difficult than high tech, precisely because it introduces interpersonal variables that might interact with the knowledge-sharing and -creation process. Can we bracket these interpersonal variables off and still get the same quality of learning? What if we take a more differentiated look at what needs to be learned? While certainly the art of cooking requires sustained direct interaction of one master and a handful of apprentices, perhaps other areas require less direct contact; perhaps direct contact may even be counterproductive.

Most knowledge-sharing platforms try to substitute direct interaction with some form of technology-enabled interaction. The reason is that it is simply more cost effective to get Mrs. Brown from the office in lower Manhattan to talk to Mr. Mueller in the Milan office by e-mail, telephone, or videoconference than to have them meet in the United States or Italy.

But is there an argument for high tech *beyond* the cost-efficiency idea? Have you ever phoned someone, hoping to leave a message rather than having to speak to them? Perhaps it is late at night over there and you do not want to disturb them, or perhaps you just do not have the inclination for a long call. You intentionally call the person at lunchtime or in the office after work[2]. Perhaps, on some occasions, technology, precisely because it severs the ties between time and space, enables us to be more purposeful in the choice of our communication media. Perhaps this purposefulness enables us to learn better and, yes, more efficiently in a high-tech environment than under high-touch circumstances.

Admittedly, this may not work in the example of the master chef and his or her apprentice. But there are other forms of learning and knowledge creation. Consider the open-source approach. Here, not one master and one apprentice interact, but tens of thousands of masters and tens of thousands of apprentices—and it is often not clear who is who and which is which. Almost all members in an open-source context have never met and never will meet. And yet, open-source development is extremely successful in the context of software development. But how does this open-source model apply to other contexts, other industries, and countries other than the United States?

CONCLUSION

When members from different organizations come together to exchange insights, share knowledge, and create value, they come together in what we call a learning network. Learning networks differ from other forms of value creation and appropriation in that they are inter-organizational, membership is not subject to formal governance processes, and they tend to involve strong elements of virtual interaction.

The concept of inter-organizational knowledge exchange is not new, having been practiced at least since the Middle Ages, where, for example, guilds provided members with learning arenas for the exchange of best practices, trade regulations, and tariffs. What is new, however, is the predominance

of virtual vs real interaction, the focus on collaboration rather than on competition, and the emphasis on value creation on the network rather than individual level.

Each of these three main processes that distinguish learning networks from their predecessors poses a key challenge, which we summarized here as (a) high tech vs high touch, (b) joint value creation vs focused value appropriation, and (c) collaboration vs competition. More research will be necessary to address each of the three key challenges identified here. In the future, we expect to see relatively more work done on the third key challenge since the role of information technology as an enabler or constraint to learning is not clear.

REFERENCES

Angehrn, A., Gibbert, M., & Nicolopoulou, K. (2003). Understanding learning networks (Guest editors' introduction). *European Management Journal, 25*(1), 559-564.

Bardaracco, J. (1991). *The knowledge link: How firms compete through strategic alliances.* Boston: Harvard Business School Press.

Davenport, T. H., & Probst, G. J. B. (2002). *Knowledge management case book* (2nd ed.). New York: John Wiley & Sons.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge.* Boston: Harvard Business School Press.

Doz, Y. (1996). The evolution of cooperation in strategic alliances. *Strategic Management Journal, 17,* 55-84.

Drucker, P. (1993). *Post capitalist society.* Oxford, UK: Butterworth-Heinemann.

Drucker, P. (1994). The age of social transformation. *The Atlantic Monthly, 27*(5), 53-80.

Dyer, J., & Nobeoka, K. (2000). Creating and managing a high performance knowledge sharing network: The Toyota case. *Strategic Management Journal,* 345-367.

Dyer, J. H., & Singh, H. (1998). The relational view: Cooperative strategy and sources of interorganizational competitive advantage. *Academy of Management Review, 23*(4), 660-679.

Gibbert, M., Angehrn, A., & Durand, T. (in press). *Learning networks: The inter-organizational side of knowledge management* (Strategic Management Society book series). Oxford, UK: Blackwell.

Grant, R. (1996). Toward a knowledge based theory of the firm. *Strategic Management Journal, 17,* 109-123.

Leibold, M., Probst, G., & Gibbert, M. (2002). *Strategic management in the knowledge economy.* Weinheim, Germany: John Wiley and Sons.

Liebeskind, J. P., Oliver, A. L., Zucker, L., & Brewer, M. (1996). Social networks, learning, and flexibility: Sourcing scientific knowledge in new biotechnology firms. *Organization Science, 7*(4), 428-443.

Polanyi, M. (1958). *Personal knowledge.* Chicago: University of Chicago Press.

Porter, M. (1980). *Competitive strategy*. New York: Free Press.

Powell, W. W. (1998). Learning from collaboration: Knowledge and networks in the biotechnology and pharmaceutical industries. *California Management Review, 40*(3), 228-240.

Powell, W. W., Koput, K. W., & Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly, 41*(1), 116-145.

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization.* New York: Doubleday/Century Business.

Spender, J. (1996a). Competitive advantage from tacit knowledge? In B. Moingeon & A. Edmondson (Eds.), *Organizational learning and competitive advantage* (pp. 56-73). London: Sage.

Spender, J. (1996b). Making knowledge the basis of a dynamic theory of the firm. *Strategic Management Journal, 17,* 45-62.

Stewart, T. A. (1998). *Intellectual capital: The new wealth of organizations.* London: Nicholas Brealey Publishing.

von Krogh, G., & Roos, J. (1995). *Organisational epistemology.* London: MacMillan.

Zuboff, S. (1988). *In the age of the smart machine: The future of work and power.* New York: Basic Books.

KEY TERMS

Collaborative Thinking: A strategic mind set where adjacent and even overlapping stages in the industry value chain are seen as potential partners rather than competitors. For example, Starbucks and Nestle both produce coffee, but they nevertheless partnered up for the creation and distribution of chilled coffee-based soft drinks, thereby leveraging Starbucks' premium brand name and Nestle's manufacturing and distribution know-how.

Community of Practice: A formal network of employees of one organization, often for the purpose of exchanging project-specific knowledge, for example, DaimlerChrysler Tech Clubs (Leibold et al., 2002).

Competitive Thinking: A strategic mind set where "us against them" prevails, and where competitive advantage denotes being better than the immediate competitor (e.g., Porter, 1980).

Knowledge Management: Managerial tools and processes geared to keep organizations from "re-inventing the wheel" by appropriate reutilization of already existing knowledge, for example, Siemens' ShareNet (Davenport & Probst, 2002).

Learning Network: An informal association of members of different organizations for the purpose of knowledge exchange. Learning networks are characterized by voluntary membership, intrinsic motivation to participate, and a focus on collaborative rather than competitive thinking.

ENDNOTE

[1] This project sought to understand better the nature of self-learning processes at work in the context of interorganizational networks, ranging from initiatives driven by local industry clusters and associations addressing relevant management- and business-related issues, to new forms of organizations of globally distributed knowledge workers operating within an open source. The project was done under the auspices of the EU IST (European Union Information Society Technologies) project Knowlaboration (Angehrn, Gibbert, & Nicolopoulou, 2003).

[2] Thanks to Barry Nalebuff for providing this example.

Learning through Business Games

Luigi Proserpio
Bocconi University, Italy

Massimo Magni
Bocconi University, Italy

BUSINESS GAMES: A NEW LEARNING TOOL

Managerial business games, defined as interactive computer-based simulations for managerial education, can be considered as a relatively new tool for adults' learning. If compared with paper-based case histories, they could be less consolidated in terms of design methodologies, usage suggestions, and results measurement.

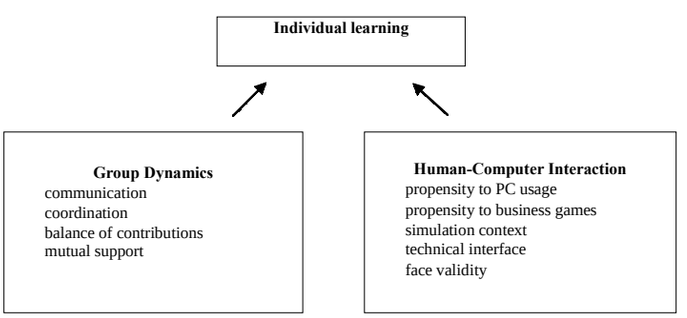
Due to the growing interest around Virtual Learning Environment (VLE), we are facing a positive trend in the adoption of business games for undergraduate and graduate education. This process can be traced back to two main factors. On the one hand, there is an increasing request for non-traditional education, side by side with an educational model based on class teaching (Alavi & Leidner, 2002). On the other hand, the rapid development of information technologies has made available specific technologies built around learning development needs (Webster & Hackley, 1997). Despite the increased interest generated by business games, many calls have still to be addressed on the design and utilization side. This contribution describes two fundamental aspects related with business games in graduate and undergraduate education: group dynamics (as current business games are almost in all instances played in groups) and human-computer interaction.

Figure 1 represents the variables that could influence individual learning in a business game context.

THE INFLUENCE OF GROUP DYNAMICS

It is widely accepted that a positive climate among subjects is fundamental to enhance the productivity of the learning process (Alavi, Wheeler, & Valacich, 1995). This is why group dynamics are believed to have a strong impact on learning within a team based context. A clear explanation of group dynamics impact on performance and learning is well developed in the teamwork quality construct (TWQ) (Hoegl & Gemuenden, 2001). Group relational dynamics are even more important when the group is asked to solve tasks requiring information exchange and social interaction (Gladstein, 1984), such as a business game. In fact, the impact of social relations is deeper when the

Figure 1. Variables influencing individual learning in a business game setting



Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

task is complex and characterized by sequential or reciprocal interdependencies among members.

With reference to TWQ, it is possible to point out different group dynamics variables with a strong influence on individual learning in a business game environment: communication, coordination, balance of contributions, and mutual support. Instructors and business games designers should carefully consider the following variables, in order to maximize learning outcomes.

Hereafter, focusing on a business game setting, we will discuss each of these concepts and their relative impact on individual learning.

Communication

In order to develop effective group decision processes, information exchange among members should also be effective. In fact, communication is the way by which members exchange information (Pinto & Pinto, 1990), and smooth group functioning depends on communication easiness and efficacy among members (Shaw, 1981).

Moreover, individuals should be granted an environment where communication is open. A lack of openness should negatively influence the integration of knowledge and group members' experiences (Gladstein, 1984; Pinto & Pinto, 1990). These statements are confirmed by several empirical studies, showing direct and strong correlation between communication and group performance (Griffin & Hauser, 1992). According to Kolb's experiential learning theory, in a learning setting based on experiential methods (i.e., business game), it is important to provide the classroom with an in-depth debriefing in order to better understand the link between the simulation and the related theoretical assumption.

For these reason, groups with good communication dynamics tend to adopt a more participative behavior during the debriefing session, with higher quality observations. As a consequence, there is a process improvement in the acquisition, generation, analysis and elaboration of information among members (Proserpio & Magni, 2004).

Balance of Contributions

It can be defined as the level of participation of each member in the group decision process. Each member, during the decision process, brings to the group a set of knowledge and experiences that allows the group to develop a cognitive advantage over individual decision process. Thus, it is necessary that each member brings his/her contributions to the group (Seers, Petty, & Cashman, 1995) in order to improve performance, learning and satisfaction of team members (Seers, 1989). A business game setting requires a good planning and implementation of strategies in order to better face the action-reaction process with the computer. For this reason, a balanced contribution among members favors the cross fertilization and the development of effective game strategies.

Coordination

A group could be seen as a complex entity integrating the various competencies required to solve a complex task. For this reason, a good balance of members' contribution is a necessary condition, although not sufficient. The expression of the group cognitive advantage is strictly tied to the harmony and synchronicity of members' contribution, that is, the degree according to which they coordinate their individual activities (Tannenbaum, Beard, & Salas, 1992).

As for communication, individuals belonging to groups with a better coordination level show better interventions in the debriefing phases. They also offer good hints to deepen the topics included in the simulation, playing as an intellectual stimulus for each other.

Mutual Support

It can be defined as the emergence of cooperative and mutually supporting behaviors, which lead to better team effectiveness (Tjosvold, 1984). In contrast, it is important to underline that competitive behaviors within a team determine distrust and frustration.

Mutual support among participants in a business game environment could be seen as an interference between the single user and the simulation: every discussion among users on simulation interpretation distracts participants from the ongoing simulation. This is why the emergence of cooperative behaviors does not univocally lead to more effective learning processes. These relations lower users' concentration and result in obstacles in the goal achievement path.

Moreover, during a business game, users play in a time pressure setting, which brings to a drop in the effectiveness of the decision process. All these issues, according to group effectiveness theories, help to understand how mutual support in a computer simulation environment could show a controversial impact on individual learning (Proserpio & Magni, 2004).

THE INFLUENCE OF HUMAN-COMPUTER INTERACTION

Business games are often described as proficient learning tools. Despite the potentiality, as stressed by Eggleston and Janson (1997), there is the need for an in-depth analysis of the relationship between user and computer. On the design side, *naïve* business games (not designed by professionals) can hinder the global performance of a simulation and bring to negative effects on the learning side. For these reason, technological facets are considered as a fundamental issue for a proficient relationship between user and computer in order to improve learning process effectiveness (Alavi & Leidner, 2002; Leidner & Jarvenpaa, 1995).

Propensity to PC Usage

Attitude toward PC usage can be defined as the user's overall affective reaction when using a PC (Venkatesh et al., 2003). Propensity to PC usage can be traced back to the concepts of pleasure, joy, interest associated with technology usage (Compeau, Higgins, & Huff, 1999). It is consistent to think that users' attitude towards computer use could influence their use involvement, increasing or decreasing the impact of simulation on learning process.

From another standpoint, more related to HCI theories, computer attitude is tied to the simulation easiness of use. It is possible to argue that a simple simulation does not require strong computer attitude to enhance the leaning process. On the contrary, a complex simulation could worsen individual learning, because the cognitive effort of the participant can be deviated from the underlying theories to a cumbersome interface.

Propensity to Business Game Usage

This construct can be defined as the cognitive and affective elements that bring the user to assume positive/negative behaviors toward a business game. In fact, in these situations, users can develop feelings of joy, elation, pleasure, depression, or displeasure, which have an impact on the effectiveness of their learning process (Taylor & Todd 1995). Consistently with Kolb's theory (Kolb, 1984) on individual different learning styles, propensity to simulations could represent a very powerful element to explain individual learning.

Simulation Context

The simulation context can be traced back to the role assumed by individuals during the simulation. In particular, it is referred to the role of participants, teacher and their relationship. Theory and practice point out that business games have to be self-explaining. In other words, the intervention of other users or the explanations of a teacher to clarify simulation dynamics have to be limited. Otherwise, the user's effort to understand the technical and interface features of the simulation could have a negative influence on learning objectives (Whicker & Sigelman, 1991). Comparing this situation with a traditional paper-based case study, it is possible to argue that good instructions and quick suggestions during a paper-based case history analysis can help in generating users' commitment and learning. On the contrary, in a business game setting, a self explanatory simulation could bring users to consider the intervention of the teacher as an interruption rather than a suggestion. Thus, simulations have often an impact on the learning process through the reception step (Alavi & Leidner, 2002), meaning that teacher's or other members' intervention hinder participants to understand incoming information.

Technical Interface

Technical interface can be defined as the way in which information is presented on the screen (Lindgaard, 1994). In a business game, the interface concept is also related to the interactivity facet (Webster & Hackley, 1997). Several studies have pointed out the influence of technical interface on

user performance and learning (Jarvenpaa, 1989; Todd & Benbasat, 1991). During the business game design, it is important to pay attention to the technical interface. It is essential that the interface captures user attention, thereby increasing the level of participation and involvement. According to the above mentioned studies, it is possible to argue that an attractive interface could represent one of the main variables that influence the learning process in a business game setting.

Face Validity

Face validity defines the coherence of simulation behaviors in relation with the user's expectancies on perceived realism. It is also possible to point out that the perceived soundness of the simulation is a primary concept concerning the users' learning (Whicker & Sigelman, 1991). The simulation cannot react randomly to the user's stimulus, but it should recreate a certain logic path which starts from player action and finishes with the simulation reaction. It is consistent with HCI and learning theories, to argue that an effective business game has to be designed to allow users to recognize a strong coherence among simulation reactions, their actions, and their behavior expectancies.

ADDITIONAL ISSUES TO DESIGN A BUSINESS GAME

The main aspect that has to be considered when designing a business game is the ability of the simulation to create a safe test bed to learn management practices and concepts. It is fundamental that users are allowed to experiment behaviors related to theoretical concepts without any real risk. This issue, together with aspects of fun and the creation of a group collaboration context, could be useful to significantly improve the learning level.

Thus, a good simulation is based on homomorphic assumptions. Starting from the existence of a reality with n characteristics, homomorphism is the ability to choose m (with $n>m$) characteristics of this reality in order to reduce its complexity without losing too much relevant information. For example, in a F1 simulation game, racing cars can have a different behavior on a wet or dry circuit, but they cannot have a different behavior among wet, very wet, or almost wet.

In order to minimize the negative impact on learning processes, it is important that characteristics not included in the simulation should not impact too much on the simulation realism.

CONCLUSION

Several studies have shown the importance of involvement and participation in the fields of standard face-to-face education and in distance learning environments (Webster & Hackley, 1997). This research note extends the validity of previous statements to the business game field.

The discussion above allows us to point out a relevant impact on learning of two types of variables, while using a business game: group dynamics and human-computer interaction.

From previous researches, it is possible to argue that the "game" dimension captures a strong part of participants' cognitive energies (Proserpio & Magni, 2004). The simulation should be designed in a fashion as interactive as possible. Moreover, instructors should take into account that their role is to facilitate the simulation flow, leaving the game responsibility to transmit experiences on theories and their effects.

This is possible if the simulation is easy enough to understand and use. In this case, despite the fact that the simulation is computer based, there is not the emergence of a strong need for computer proficiency. This conclusion is consistent with other researches which showed the impact of the easiness of use on individual performance and learning (Delone & McLean, 1992).

The relationship between user and machine is mediated by the interface designed for the simulation, which represents a very powerful variable to explain and favor the learning process with these high involvement learning tools.

Computer simulations seem to have their major strength in the computer interaction, which ought to be the main focus in the design phase of the game. Interaction among groups' members is still important, but less relevant than the interface on individual learning.

REFERENCES

Alavi, M. & Leidner D. (2002). Virtual learning systems. In H. Bidgole (Ed.), *Encyclopedia of Information Systems* (pp. 561-572). Academic Press.

Alavi, M., Wheeler, B.C., & Valacich, J.S. (1995). Using IT to reengineer business education: An exploratory investigation of collaborative telelearning, *MIS Quarterly, 19*(3), 293-312.

Compeau, D.R., Higgins, C.A., & Huff, S. (1999). Social cognitive theory and individual reactions to computing technology: A longitudinal study. *MIS Quarterly, 23*(2), 145-158.

Delone, W.H. & McLean, E.R. (1992). Information systems success: The quest for dependent variables. *Information Systems Research, 3*(1), 60-95.

Eggleston, R.G. & Janson, W.P. (1997). Field of view effects on a direct manipulation task in a virtual environment. *Proceedings of the Human Factors and Ergonomic Society 41st Annual Meeting*, (pp. 1244-1248).

Gladstein, D.L. (1984). Groups in context: A model of task group effectiveness. *Administrative Science Quarterly, 29*, 499-517.

Griffin, A. & Hauser, J.R. (1992). Patterns of communication among marketing, engineering, and manufacturing: A comparison between two new product development teams. *Management Science, 38*(3), 360-373.

Hoegl, M. & Gemuenden, H.G. (2001). Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science, 12*(4), 435-449.

Institute of Electrical and Electronics Engineers (1990). *IEEE standard computer dictionary: A compilation of IEEE standard computer glossaries.* New York.

Jarvenpaa, S.L. (1989). The effect of task demands and graphical format on information processing strategies. *Management Science, 35*(3), 285-303.

Kolb, D.A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice-Hall.

Leidner, D.E. & Jarvenpaa, S.L. (1995). The use of information technology to enhance management school education: A theoretical view. *MIS Quarterly, 19*(3), 265-292.

Lindgaard, G. (1994). *Usability testing and system evaluation: A guide for designing useful computer systems*. London; New York: Chapman & Hall.

Pinto, M.B. & Pinto, J.K. (1990). Project team communication and cross functional cooperation in new program development. *Journal of Product Innovation Management, 7*, 200-212.

Proserpio, L. & Magni, M. (2004). To play or not to play. Building a learning environment through computer simulations. *ECIS Proceedings*, Turku, Finland.

Seers, A., Petty, M., & Cashman, J.F., (1995). Team-member exchange under team and traditional management: A naturally occurring quasi experiment. *Group & Organization Management, 20*, 18-38.

Seers, A. (1989). Team-member exchange quality: A new construct for role-making research. *Organizational Behavior and Human Decision Process, 43*, 118-135.

Shaw, M.E. (1981). *Group dynamics: The psychology of small group behavior*. New York: McGraw-Hill.

Tannenbaum, S.I., Beard, R.L., & Salas, E. (1992). Team building and its influence on team effectiveness: An examination of conceptual and empirical developments. K. Kelley, (a cura di), *Issues, Theory, and Research in Industrial/Organizational Psychology*. Elsevier, Amsterdam, Holland, 117-153.

Taylor, S. & Todd, P.A. (1995). Assessing IT usage: The role of prior experience. *MIS Quarterly, 19*(2), 561-570.

Tjosvold, D. (1984). Cooperation theory and organizations. *Human Relations, 37*(9), 743-767.

Todd, P.A. & Benbasat, I. (1991). An experimental investigation of the impact of computer based decision aids on decision making strategies. *Information Systems Research, 2*(2), 87-115.

Venkatesh, V. et al. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478.

Webster, J. & Hackley, P. (1997). Teaching effectiveness in technology-mediated distance learning, *Academy of Management Journal, 40*(6), 1282-1310.

Whicker, M.L. & Sigelman, L. (1991). *Computer simulation applications: An introduction.* Newbury Park, CA: Sage Publications.

Wight, A. (1970). Participative education and the inevitable revolution. *Journal of Creative Behavior, 4*(4), 234-282.

KEY TERMS

Business Games: Computer-based simulations designed to learn business-related concepts.

Experiential Learning: A learning model "which begins with the experience followed by reflection, discussion, analysis and evaluation of the experience" (Wight, 1970, p. 234-282).

HCI (Human-Computer Interaction): A scientific field concerning design, evaluation, and implementation of interactive computing systems for human usage.

Interface: An interface is a set of commands or menus through which a user communicates with a software program.

TWQ (Teamwork Quality): A comprehensive concept of the quality of interactions in teams. It represents how well team members collaborate or interact.

Usability: The ease with which a user can learn to operate, prepare inputs for, and interpret outputs of a system or component (Institute of Electrical and Electronics Engineers).

VLE (Virtual Learning Environments): Computer-based environments for learning purposes.

Local Loop Unbundling

Alessandro Arbore
Bocconi University, Italy

IN ESSENCE

Local loop unbundling (LLU) is one of the most important and controversial policy instruments adopted in many countries since the second half of the 1990s. Its aim is to foster competition within local telecommunication markets.

LLU requires any former monopolist (i.e., the incumbent) to lease, at cost, part of its local network facilities to any requesting competitor (i.e., the new entrants). The local assets that can be leased from the incumbent are called unbundled network elements (UNEs).

INTRODUCTION AND DEFINITIONS

A critical issue in the context of telecommunications market openness is the access to the local network (as defined in Table 1). It is critical because a local network allows telecommunication service providers to reach the end users. It is especially critical because, despite the recent liberalization of the industry, a combination of historic and structural factors grant incumbent operators a strong, privileged position.[1]

One of the regulatory answers given in recent years is the obligation, for the incumbent, to share part of its local facilities with new operators. The possibility to lease the incumbent's local network assets is generally referred to as unbundling of the local loop. As this article shows, the incumbent's legal obligations to provide such access can be more or less burdensome, from both a technical and an economic point of view.

BACKGROUND

The history of telecommunications in developed countries is the history of a monopolistic, vertically integrated industry that regulators, year after year, have tried to take back to competition. Specific technical and economic conditions (see Note 1) made and make this a tremendous challenge.

The long process toward competition started in the U.S. during the 1950s and 1960s, when the monopoly for terminal equipment—then justified with "network integrity" arguments—was first disputed.[2] Eventually, the long distance monopoly, then considered a natural monopoly,[3] was also challenged. A series of decisions in the United States (U.S.) during the 1960s and 1970s testify an increasing desire to overcome the status quo, although in a context of high uncertainty for the political and economic consequences (Brock, 1994).[4] The process accelerated during the 1980s, with the divestiture of AT&T in 1982 and, since 1984, with the duopoly policy promoted by the Thatcher Administration in the United Kingdom (U.K). With the privatization of British Telecom, the U.K. also devised new forms of "incentive regulation".[5] During the 1990s, the positive results in these pioneering countries prompted liberalization reforms worldwide.

The local telecommunications market seems to be the last bastion of the monopolistic era. Indeed, in the last decade, technological innovation and demand growth weakened the idea of a local natural monopoly (see Note 3). Accordingly, the U.S. Congress removed legal barriers to entry in 1996;[6] the European Parliament and the Council required member states to do the same by January 1998.[7] Yet, after several years, the incumbent operator still dominates local telecommunications.

THE POLICY MEASURES

Current regulations in the U.S. and European Union (EU) seek to encourage local competition by reducing entry barriers for new competitors. To that end, different rules facilitate alternative methods of entering the market. The strategy of a new entrant can be based on one, or a mix of the following methods.

First, new competitors can purchase incumbents' services on a wholesale basis and resell them under

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. Preliminary definitions

Generically, the expressions "local network", "local loop", "local access", or "access network" can be used equally to refer to all local telecommunication assets, including switching and "last mile" transport facilities. The expression "local" has a spatial meaning and typically refers to an urban area. The expression "last mile" informally refers to the part of the public switched telephone network (PSTN) that extends from the customer premises equipment (CPE) to the first network's switching center (the central office, also called local or switching exchange). In plain English, it is the physical connection – generally made of a pair of copper wires – between the subscriber's location and the nearest telephone exchange. The last mile, which is also called "line" or "subscriber line", coincides with the most restrictive definition of local loop. A "local telecommunications market" may include the provision of: - calls (voice or data) originated and terminated within a given urban area; - enhanced features such as touch-tone calling or call forwarding; - access to local services by other providers (e.g. long distance), which are charged for using the local network; - and high speed Internet access services, like DSL services and cable-modem; such that a small but significant and non-transitory increase in price (SSNIP) above the competitive level will be profitable for a hypothetical monopolist. (This integrates the definition by Harris and Kraft, 1997, and the Federal Trade Commission-Department of Justice Merger Guidelines, as included in Woroch's definition, 1998).

their own brand. Where using this strategy, a firm is said to operate as a "reseller". Regulations tend to set wholesale prices on a discount basis ("price minus" mechanism): typically, wholesale prices are set equal to the retail prices minus commercial, billing, and other avoidable costs.[8]

Second, new competitors can build their own loop or upgrade an existing local communication network (i.e. cable TV). In this case, the law grants the right to interconnect to the public telecommunications network, so that network externalities do not preclude competition.[9] When using this strategy, a firm is operating as an "infrastructure provider". The resulting competition is referred to as facility-based competition.[10] In the U.S., as in the EU, interconnection must be provided at cost, at any technically feasible point, at non-discriminatory conditions, and ensuring the same quality of the incumbent's services. The kind of costs to be accounted for varies among the countries.[11]

Third, and most important here, new entrants can provide local services by leasing specific facilities ("elements") from the incumbent's network. As said, this practice is the unbundled access to the local loop. Where using unbundled elements, a firm can be said to operate as a service provider. Service providers foster a service competition among players that actually rely on the same infrastructure. An unsolved thorny issue is which form of competition—service or facility-based—delivers the highest social returns and under which circumstances.

More details on unbundling policies in the U.S. and EU are provided in the next sections.

OVERVIEW OF THE U.S. UNBUNDLING POLICY

Section 251(c)(3) of the Telecommunications Act of 1996 decrees, for incumbent local exchange carriers (ILECs), "[t]he duty to provide, to any requesting telecommunications carrier (…) nondiscriminatory access to network elements on an unbundled basis at any technically feasible point on rates, terms, and conditions that are just, reasonable, and nondiscriminatory (…)." The controversial expression "at any technically feasible point" is blurred by section 251(d)(2): "In determining what network elements should be [unbundled], the [FCC] shall consider, *at a minimum*, whether– (A) access to such network elements (…) is *necessary*; and; (B) the failure to provide access to such network elements would *impair* the ability of the telecommunications carrier seeking access to provide the services that it seeks to

offer." (Parenthesis and italic mine. These are known as the "necessary" and "impair" requirements, as in 525 U.S. 366 [1999]).

Following section 252(d)(1), a just and reasonable rate "shall be based on the cost (determined without reference to a rate-of-return or other rate-based proceeding)," and "may include a reasonable profit." The generic expression "shall be based on the cost" left significant discretionary power to the implementers.

According to a 1999 decision of the U.S. Supreme Court, "the FCC has general jurisdiction to implement the 1996 Act's local-competition provisions."[12] The decision legitimates the FCC Order of August 1996 that established, among other things, uniform national rules for unbundling conditions (hereafter "FCC Order").[13]

In fact, the 1996 Act provides that private negotiations are the starting point for agreements between the new entrants and the ILEC (section 252). When the parties fail to reach an agreement—which is likely, given the bargaining power of the incumbent—they are entitled to ask arbitration from the appropriate state commission.[14] At that point, according to the Supreme Court decision, the state commissions should essentially administer the pricing guidelines provided by the FCC Order.

The FCC interpreted the generic pricing rule of the Congress (i.e. unbundling rates "based on the cost") to mean rates based on forward-looking long-run incremental costs. Essentially, the FCC pricing methodology (labeled TELRIC, "total element long run incremental cost") estimates the overall additional cost supported by the incumbent when a certain new element is introduced in its network, but under the hypothesis that the network is built with the most efficient technology available.[15] A "reasonable" share of forward-looking common costs can be allocated to the unbundled elements.[16]

A further critical point, other than the pricing methodology, is the identification of the elements to be unbundled. Not surprisingly, the FCC interpretations of the "necessary" and "impair" requirements of §251(d)(2) (see above) has been—and probably will be—at the core of different litigations.[17] As at the time of this writing, the commission is re-examining its unbundling framework exploring many of the issues that the courts raised.[18] The main criticism moved by authors like Harris and Kraft (1997) and Jorde, Sidak, and Teece (2000) was that the FCC interpretation did not limit mandatory unbundling to "essential" facilities.[19]

OVERVIEW OF THE EU UNBUNDLING POLICY

In July 2000, the European Commission presented its guidelines for a "New Regulatory Framework for electronic communications, infrastructure, and associated services". The guidelines followed the decisions taken in Lisbon on March 23-24 2000, when the European Council launched the "eEurope" program to foster the benefits of a "digital economy".

The guidelines in the matter of unbundling were illustrated in the "Proposal for a Regulation of the European Parliament and of the Council on unbundled access to the local loop," adopted on July 12, 2000.[20] After a wide public consultation, the proposal was converted into the Regulation EC No 2887/2000 on December 18, 2000.[21] European regulations, as opposed to European directives, are directly applicable in all member states, without the need of a national implementation.[22] Therefore, the European unbundling provisions are automatically enforced in every member state.

Before reviewing the regulation, it must be noted that the EU identifies three arrangements for unbundled local access services:[23]

- Full unbundling of the local loop: a third party rents the local loop from the incumbent for its exclusive use.
- Shared access to the local loop (also known as "spectrum sharing", "bandwidth sharing", or "line splitting"): the line is split into a higher and a lower frequency portion, allowing the lower frequency portion to be used for voice and the higher frequency portion for data transmission (typically for high-speed Internet access). A third party is then entitled to request access just to the higher portion. In this way, the incumbent continues to provide telephone services, while the new entrant delivers high-speed data services using its own high-speed modems.[24]
- High speed bit-stream access: similar to shared access, but the high-speed elements too (like

ADSL modems) are leased from the incumbent. The third party does not have actual access to the copper pair in the local loop.

The European Regulation (EC) No 2887/2000 mandates unbundled access only to the metallic local loops (copper or aluminium) and only for the operators that have been designated by their national regulatory authorities (NRA) as having "significant market power" in the fixed telephone market (so-called "notified operators").[25] Requests can only be refused for reasons of technical feasibility or network integrity (art. 3, sub2).

Article 2 specifies that "unbundled access to the local loop" means both "full unbundled access" and "shared access" to the loop.

As in the U.S., commercial negotiation is the preferred method for reaching agreement on technical and pricing issues for local loop access. Nonetheless, the intervention of the NRA is always possible in order to ensure: 1) fair competition 2) economic efficiency, 3) maximum benefit for end-users. Especially, the NRA must have the power to impose changes to a reference offer that the incumbent must publish, as well as to require from the players all the necessary information.

"Costing and pricing rules … should be transparent, non-discriminatory and objective to ensure fairness. Pricing rules should ensure that the local loop provider is able to cover its appropriate costs in this regard plus a reasonable return, in order to ensure the long term development and upgrade of local access infrastructure."[26] The regulation underlines that pricing rules must bear in mind the importance of new infrastructure investments.

Finally, the regulation specifies that member states can still "maintain or introduce measures in conformity with [European] Community law which contain more detailed provisions than those set out in this Regulation …"[27] Few additional suggestions for the national regulator are provided in the Commission Recommendation 2000/417/EC of May 25, 2000.[28] In particular, forward-looking approaches based on current costs (i.e. "the costs of building an efficient modern equivalent infrastructure today") seem to be suggested to price unbundled elements in the early stages of competition.[29]

The rules of the regulation, however, leave some discretionary powers to the member states, especially in the definition of the incumbent's costs to be accounted for (leading to higher or lower rental charges). The powers for national implementation are generally shared between a Telecommunications Ministry and the national telecommunications authority.

CONCLUSION

In recent years, telecom regulators have considered removal of legal barriers to entry in the local telecommunications industry as insufficient, alone, to start an effective competitive process. Local loop unbundling is one of the most important and controversial policy instrument designed by the regulators to foster competition in these markets. Two final considerations can be made.

First, it is important to keep in mind that competition should not be considered as the ultimate goal for a regulator. Instead, among its goals there is economic efficiency. Competition ensures economic efficiency only in the absence of market and government failures. This is not the case with local telecommunications, unfortunately: a unique combination of historic and structural peculiarities mentioned in this work, in fact, may prevent free-market forces from leading to the most efficient allocation of resources in the industry.

From this, it follows the second observation: beyond the support that LLU may provide to competition, some commentators hypothesize that its current implementation—especially in the U.S.—might negatively affect innovation, investment, and product development for both the incumbents and the new entrants. In the long run, it is argued, the overall result might be a lower level of economic efficiency.

Although at this time there is no clear-cut evidence of such detrimental effects, it is probably necessary to perform a more comprehensive assessment of the policy's net social benefits. The need for this information appears especially pressing because there are signals of high implementing costs in front of moderate results.

REFERENCES

Bauer, J.M. (1997). Market power, innovation and efficiency in telecommunications: Schumpeter reconsidered. *Journal of Economic Issues,* (2), 557-565.

Baumol, W.J. (1983). Some subtle pricing issues in railroad regulation. *International Journal of Transportation Economics, 10*, 341-355.

Baumol, W.J. & Sidak, J.G. (1994). *Toward competition in local telephony*. Cambridge, MA: MIT Press.

Brock, G.B. (1994). *Telecommunication policy for the information age*. Cambridge, MA: Harvard University Press.

Brock, G.B. & Katz, M.L. (1997). Regulation to promote competition: A first look at the FCC's implementation of the local competition provisions of the telecommunications act of 1996. *Information Economics and Policy,* (2), 103-117.

DSTI-ICCP-TISP(2000)3-FINAL

Economides, N. (2000). Real options and the costs of the local telecommunications network. In J. Alleman & E. Noam (Eds.), *The new investment theory of real options and its implications for cost models in telecommunications*. Boston: Kluwer Academic Publishers.

Economides, N. & Flyer, F. (1997). *Compatibility and market structure for network goods*. (Stern School of Business, NYU, Discussion Paper EC-98-02). [Electronic version]. Retrieved September 13, 1999, from: *http://raven.stern.nyu.edu/networks/98-02.pdf*

Faulhaber, G.R. & Hogendorn, C. (2000). The market structure and broadband telecommunications. *The Journal of Industrial Economics,* (3), 305-329.

Federal Communication Commission (FCC) (1996). *Implementation of the local competition provisions in the Telecommunications Act of 1996*. (CC Docket No. 96-98, FCC 96-325). Retrieved March 9, 2002, from: *http://www.fcc.gov/ccb/local_competition/fcc96325.pdf*

Greenstein, S., McMaster, S. & Spiller, P. (1995). The effect of incentive regulation on infrastructure modernization: Local companies' deployment of digital technology. *Journal of Economics and Management Strategy,* (2), 187-236.

Harris, R. & Kraft, C.K. (1997). Meddling through: Regulating local telephone competition in the United States. *Journal of Economic Perspectives, 11*(4), 93-113.

Hausman, J.A (1997). Valuing the effect of regulation on new services in telecommunications. *Brookings Papers on Economic Activity, Microeconomics*, 1-38.

Jorde, M., Sidak, G.J. & Teece, D.J. (2000). Innovation, investment, and unbundling. *Yale Journal on Regulation,* (1), 1-36.

Kahn, A.E. (1998). *Letting go: Deregulating the process of deregulation, or temptation of the kleptocrats and the political economy of regulatory disingenuousness*. East Lansing, MI: Michigan State University.

Katz, M.L. & Shapiro, C. (1994). Systems competition and network effects. *Journal of Economic Perspectives, 8*, 93-115.

Kiessling, T. & Blondeel, Y. (1999). The impact of regulation on facility-based competition in telecommunications. *Communications & Strategies, 34,* 19-44.

Laffont, J.J. & Tirole, J. (1991). The politics of government decision-making: A theory of regulatory capture. *Quarterly Journal of Economics, 106*, 1089-1127.

Laffont, J.J. & Tirole, J. (2000). *Competition in telecommunications*. Cambridge, MA: The MIT Press.

Liebowitz, S.J. & Margolis, S.E. (1995). Are network externalities a new source of market failure? *Research in Law and Economics, 17*, 1-22.

Majumdar, S.K. & Chang, H.H. (1998). Optimal local exchange carrier size. *Review of Industrial Organization,* (6). 637-649.

Mason, R. & Valletti, T.M. (2001). Competition in communication networks: Pricing and regulation. *Oxford Review of Economic Policy,* (3), 389-415.

Organisation for Economic Co-Operation and Development (OECD) (2001). *Interconnection and local competition*. (Working paper DSTI/ICCP/TISP(2000)3/FINAL). [Electronic version]. Retrieved December 11, 2001, from *http://www.olis.oecd.org/olis/2000doc.nsf/LinkTo/*

Organisation for Economic Co-Operation and Development (OECD) (1996). *The essential facilities concept.* (OCDE/GD(96)113). [Electronic version]. Retrieved August 15, 2001, from *http://www1.oecd.org/daf/clp/roundtables/ESSEN.PDF*

Ros, A. (1998, June). *Does ownership or competition matter? The effects of telecommunications reform on network expansion and efficiency.* Paper presented at the 12th Biennial Conference of the International Telecommunications Society, Stockholm, Sweden.

Rosston, G.L. (1997). Valuing the effect of regulation on new services in telecommunications. *Brookings Papers on Economic Activity, Microeconomics*, 48-54.

Roycroft, T.R. (1998). Ma Bell's legacy: Time for a second divestiture? *Public Utility Fortnightly,* (12), 30-34.

Shin, R.T. & Ying, J.S. (1992). Unnatural monopolies in local telephone. *Rand Journal of Economics,* (2), 171-183.

Sidak, J.G. & Spulber, D.F. (1997a). *Deregulatory takings and the regulatory contract: The competitive transformation of network industries in the United States*. Cambridge: Cambridge University Press.

Sidak, J.G. & Spulber, D.F. (1997b). Givings, takings, and the fallacy of forward-looking costs. *New York University Law Review*, *72*, 1068-1164.

Sidak, J.G. & Spulber, D.F. (1997c). The tragedy of the telecommons: Government pricing of unbundled network elements under the Telecommunications Act of 1996. *Columbia University Law Review*, *97*, 1201-1281.

Taschdjian, M. (1997, September). *Alternative models of Telecommunications policy: Service competition versus infrastructure competition*. Paper presented at the 25th Annual Telecommunications Policy Research Conference, Alexandria, VA.

Tomlinson, R. (1995). The impact of local competition on network quality. In W. Lehr (Ed.), *Quality and reliability of telecommunications infrastructure.* Mahwah, NJ: Lawrence Erlbaum Associates.

Vogelsang, I. & Mitchell. B. (1997). *Telecommunications competition: the last ten miles*. Cambridge, MA: The MIT Press.

Willig, R.D. (1979). The theory of network access pricing. In H.M. Trabing (Ed.), *Issues in public utility regulation* (pp. 109-152). East Lansing, MI: Michigan State University.

Woroch, G.A. (1998). *Facilities competition and local network investment: theory, evidence and policy implications*. (University of California at Berkeley, Working Paper CRTP-47). [Electronic version]. Retrieved June 2, 2001, from *http://groups.haas.berkeley.edu/imio/crtp/publications/workingpapers/wp47.PDF*

Zolnierek, J., Eisner J., & Burton E. (2001). An empirical examination of entry patterns in local telephone markets. *Journal of Regulatory Economics, 19*, 143-160.

KEY TERMS

Access Network: See "local network".

Forward-Looking Long-Run Incremental Costs: See "TELRIC."

Incentive Regulation: Simply stated, it refers to a variety of regulatory approaches (starting with "price caps") that attempt to provide or enhance incentives for utilities to operate more efficiently. Incentive regulation is a response to the limits of the traditional "rate of return regulation," which set rates so as to cover operating expenses and ensure a "reasonable" return on invested capital. This was administratively cumbersome, detrimental to efficiency, and subject to the risk of overcapitalizations.

Last Mile: Informally refers to the part of the public switched telephone network (PSTN) that

extends from the customer premises equipment (CPE) to the first network's switching center (the central office, also called local or switching exchange). In plain English, it is the physical connection—generally made of a pair of copper wires—between the subscriber's location and the nearest telephone exchange. The last mile, which is also called "line" or "subscriber line", coincides with the most restrictive definition of local loop.

Line: See "last mile".

Local Access: See "local network".

Local Loop: See "local network". See also, for a more restrictive definition, "last mile".

Local Loop Unbundling (LLU): One of the most important and controversial policy instrument adopted in many countries since the second half of the 1990s to foster the competitive process in local telecommunication markets. LLU codifies the legal obligation for the incumbent operator to provide, at cost, part of its local network facilities (unbundled elements) to its competitors.

Local Network: Refers to all local telecommunication assets, including switching and last mile transport facilities. The expression "local" has a spatial meaning and typically refers to an urban area.

Local Telecommunications Market: Generally include the provision of: calls (voice or data) originated and terminated within a given urban area; enhanced features such as touch-tone calling or call forwarding; access to local services by other providers (e.g., long distance), which are charged for using the local network; and high speed Internet access services, like DSL services and cable-modem; such that a small but significant and non-transitory increase in price (SSNIP) above the competitive level will be profitable for a hypothetical monopolist. (This integrates the definition by Harris and Kraft, 1997, and the Federal Trade Commission-Department of Justice Merger Guidelines, as included in Woroch's definition, 1998).

Natural Monopoly: Simply stated, economists refer to a natural monopoly when very high fixed costs are such that it would be inefficient for more than one firm to supply the market because of the duplication in fixed costs involved. More formally, this means that long run marginal costs are always below long run average costs.

Subscriber Line: See "last mile".

TELRIC: Total Element Long Run Incremental Cost. It is the FCC pricing methodology for local loop unbundling. It is based on forward-looking long-run incremental costs: essentially, the regulator estimates the overall additional cost supported by the incumbent when a certain new element is introduced in its network, but under the hypothesis that the network is built with the most efficient technology available.

ENDNOTES

1 Among these factors, it is possible to recall issues like natural monopoly, network externalities, and universal service, introduced later in this article and deepened in different parts of this encyclopedia.

2 See, in particular, the Hush-A-Phone case in 1956 (238 F.2d 266, D.C. Cir. 1956) and the Carterfone case in 1968 (14 FCC 2d 571, 1968).

3 Simply stated, we refer to a natural monopoly when very high fixed costs are such that it would be inefficient for more than one firm to supply the market because of the duplication in fixed costs involved. More formally, this means that long run marginal costs are always below long run average costs. Marshall (1927), Baumol (1977), and Sharkey (1982) establish today's dominant theory for identifying a natural monopoly.

4 See, in particular, the FCC "Above 890" decision (27 F.C.C. 359, 1959) and the "Specialized Common Carrier" decision in 1971 (29 FCC 2d 870, 1971).

5 Simply stated, incentive regulation refers to a variety of regulatory approaches (starting with "price caps") that attempt to provide or enhance incentives for utilities to operate more efficiently. Incentive regulation is a response to the limits of the traditional "rate of return regulation", which set rates so as to cover operating expenses and ensure a "reasonable" return on invested capital. This was administratively cumbersome, detrimental to efficiency, and subject to the risk of overcapitalizations.

[6] Telecommunications Act of 1996, Pub. LA. No. 104-104, 110 Stat. 56 (1996). Codified at 47 U.S.C. 151 et Seq. In the following, where not mentioned otherwise, it is referred as "the 1996 Act".

[7] Directive 96/19/EC amending Directive 90/388/EEC with regard to the implementation of full competition in telecommunications markets (OJ No L 74, 22.3.96).

[8] For the U.S., see 47 U.S.C. § 251(c)(4) and § 252(d)(3). The theoretical rationale for "price minus" mechanisms lays in the efficient component pricing rule (ECPR), also known as the Baumol-Willig rule or party pricing principle. For more information on ECPR: Baumol (1983), Willig (1979), Baumol and Sidak (1994), Sidak and Spulber (1997a).

[9] Network externalities exist because the value of telephone services to a subscriber increases as the number of subscribers grows. A new subscriber, then, derives private benefits, but also confers external benefits on the existing subscribers: they are now able to communicate with him. The important consequence in the context of market openness is that a new network, starting with zero subscribers, would have no chance to compete with the incumbent network.

[10] The literature sometimes split facility-based competition into "inter-modal facility-based competition" (competition among networks using different technologies, generally different transmission media) and "intra-modal facility-based competition" (competition for carrying services among networks using the same technology).

[11] In the U.S., the 1996 Act provides that "charges for transport and termination of traffic" should be priced considering the "additional costs" incurred by the incumbent (§ 252(d)(2)(ii)). The Federal Communication Commission (FCC) interpreted this expression with a pricing methodology based on forward-looking long-run incremental costs (CC Docket No.96-98, *Implementation of the Local Competition Provisions in the Telecommunications Act of 1996*, FCC 96-325, released August 8, 1996).

[12] AT&T Corp. v. Iowa Utilities Bd., 525 U.S. 366 (1999).

[13] CC Docket No.96-98, Implementation of the Local Competition *Provisions in the Telecommunications Act of 1996*, FCC 96-325, released August 8, 1996.

[14] Because of the "non-discriminatory" obligations only the first agreements within a contractual period (round) tend to require the intervention of the state commission: the following ones can simply ask for the same conditions of the most favorable agreements.

[15] For critiques on this choice, see among the others, Sidak and Spulber (1997b, 1997c), Hausman (1997), and Harris and Kraft (1997).

[16] FCC Order, 672-702 (the incumbent bears the burden of proving the existence of common costs).

[17] *Iowa Utils. Bd. v. FCC*, 120 F.3d 753 (8th Cir. 1997); and *U.S. Telecom Ass'n v. FCC*, 290 F.3d 415 (D.C. Cir. 2002).

[18] For the FCC interpretation as on February 2003, see Docket No.CC 01-338 (February 20, 2003) at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-231344A2. In a first approximation, mandatory unbundling is not required for ILEC network elements that are deployed to provide broadband access, but for traditional voice related network elements only.

[19] "The European Commission defines 'essential' any facility or infrastructure without access to which competitors cannot provide services to their customers (*Sea Containers vs Stena Sealink* [OJ, 18/1/94, L15 p.8]). (…) In the US definition, it is quite clear that a facility is essential only if the services it can provide belong to a relevant market (no duplicability at reasonable costs) controlled by a monopolist" (OECD, 1996, p. 55).

[20] COM(2000) 394, July 12, 2000.

[21] Official Journal L 336, 30/12/2000 p.4-8. In the following, where not mentioned otherwise, it is referred to as "the Regulation".

[22] Art.249 of the EC Treaty, subs.2. Directives, on the other hand, are less stringent and have to be transformed by the Member states into national law (art. 249 of the EC Treaty, subs.3).

[23] Commission of The European Communities, Commission Recommendation "On unbundled Access to the Local Loop," Brussels 26th April C(2000)1059.

[24] A mandatory unbundling of the higher frequency portion of the line has been considered also by some U.S. States, like California (Jorde et al., 2000). In its latest orientation, however, the FCC no longer requires that line-sharing be available as an unbundled element (FCC, Docket No. CC 01-338, February 20, 2003).

[25] For the determination of significant market power, then, the "relevant market" is not the local telecommunications market, but the entire fixed telephone market (in the national geographical area within which an organization is authorized to operate). Within the original European regulatory framework, "significant market power [was] not the same as the concept of dominant position used in competition law. Significant market power [was] an [Open Network Provision] concept used to decide when an organization should [have been] subject to specific obligations (…). An organization [was] presumed to have significant market power if it [had] more than 25% of the relevant market." (European Commission, DG XIII, *Determination of Organizations with Significant Market Power (SMP) for implementation of the ONP Directives*, p.3, Brussels, 1st March 1999. Parentheses mine). Recently, this approach has been revised and more closely oriented to the competition law principles. Accordingly, "the threshold for imposing ex-ante obligations – new SMP – is now aligned to the competition law concept of dominance (i.e. the power of an undertaking, either alone or jointly with others, to behave to an appreciable extent independently of competitors, consumers and ultimately consumers)" (ITU, World Telecommunication Development Conference (WTDC-02), INF-24 E, p.3, Istanbul, March 23, 2002).

[26] Regulation (EC) No 2887/2000, Sub. 11 of the preamble.

[27] *Ib.* art. 1, sub 4.

[28] Official Journal L 156, 29/06/2000 p.44 – 48.

[29] *Ib.* art 1, sub. 6.

Local Loop Unbundling Measures and Policies in the European Union

Ioannis P. Chochliouros
Hellenic Telecommunications Organization S.A. (OTE), Greece

Anastasia S. Spiliopoulou-Chochliourou
Hellenic Telecommunications Organization S.A. (OTE), Greece

George K. Lalopoulos
Hellenic Telecommunications Organization S.A. (OTE), Greece

INTRODUCTORY FRAMEWORK: THE CHALLENGE

Recent European policies have very early identified (European Commission, 1999) the great challenge for the European Union (EU) to promote various liberalisation and harmonisation measures in the relevant electronic communications markets to support initiatives for competition, innovation, development, and growth (Chochliouros & Spiliopoulou-Chochliourou, 2003). In order to fully seize the growth and job potential of the digital, knowledge-based economy, it has been decided that businesses and citizens should have access to an inexpensive, world-class communications infrastructure and a wide range of modern services, especially to support "broadband" evolution and multimedia penetration. Moreover, different means of access must prevent information exclusion, while information technologies should be used to renew urban and regional development and to promote environmentally sound technologies. A fundamental policy was to introduce greater competition in local access networks and support local loop unbundling (LLU) in order to help bring about a substantial reduction in the costs of using the Internet and to promote high-speed and "always-on" access.

The "local loop" mainly refers to the physical copper-line circuit in the local access network connecting the customer's premises to the operator's local switch, concentrator, or equivalent facility. Traditionally, it takes the form of twisted metallic pairs of copper wires (one pair per ordinary telephone line); fiber-optic cables are being deployed increasingly to connect large customers, while other technologies are also being rolled out in local access networks (such as wireless and satellite local loops, power-line networks, or cable TV networks). Although technology's evolution and market development are very rapid, the above alternatives—even in a combined use—cannot provide adequate guarantees to ensure the sufficient and nationwide spreading of LLU in a reasonable time period and to address the same customer population, if practically compared to the digital subscriber loop (DSL) option, offered via the existing copper. Until very recently, the local access network remained one of the least competitive segments of the liberalised European telecommunications market (European Commission, 2001) because new entrants did not have widespread alternative network infrastructures and were "unable" with traditional technologies to match the economies of scale and scope of operators notified as having significant market power (SMP) in the fixed network (European Parliament & European Council, 1997). This resulted from the fact that operators rolled out their old copper local access networks over significant periods of time, protected by exclusive rights, and were able to fund their investment costs through monopoly rents. However, a great challenge exists as the Internet-access market is rapidly becoming a utility market. Prices for customer premises equipment (CPE) are based on commodity product pricing, while digital subscriber-line services are beginning to be considered by the consumer as a utility service in the same view as the telephone or electricity network.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

THE AIM OF THE RECENT EUROPEAN POLICIES: TOWARD AN INNOVATIVE FUTURE

The importance to new entrants of obtaining unbundled access to the local loop of the fixed incumbent across the EU (and the entire European Economic Area [EEA]) was strongly acknowledged by the European Commission, which has promoted early initiatives in this area, in particular, with its adoption in April 2000 of a recommendation (European Commission, 2000b) and then an associated communication (European Commission, 2000a) on LLU. These measures were reinforced by the announcement that a legally binding provision for unbundling would be included in the new regulatory framework (Chochliouros & Spiliopoulou-Chochliourou, 2003).

The basic philosophy of the proposed measures to liberalise the markets was the estimation that it would not be economically viable for new entrants to duplicate the incumbent's copper local loop and access infrastructure in its entirety and in a reasonable time period, while any other alternative infrastructures (e.g., cable television, satellite, wireless local loops) do not generally offer the same functionality or ubiquity.

LLU has a large impact on both the deployment rules and the engineering of broadband systems (Ödling, Mayr, & Palm, 2000). The motivation for liberalising the European telecommunications market was to increase competition and, consequently, to provide faster development of services and more attractive tariffs. In order to achieve the projected target, and following the regulatory practices already applied in the United States, the European Commission obliged operators having SMP in the fixed network to unbundle their copper local telecommunications loop by December 31, 2000. This was, in fact, a first measure to promote the opening of the local markets to full competition and the introduction of enhanced electronic communications. The related argumentation was based on the fact that existing operators could roll out their own broadband, high-speed bit-stream services for Internet access in their copper loops, but they might delay the introduction of some types of DSL technologies and services in the local loop where these could substitute for the operator's current offerings. Any such delays would be at the expense of the end users; therefore, it was appropriate to allow third parties to have unbundled access to the local loop of the SMP (or "notified") operator, in particular, to meet users' needs for the competitive provision of leased lines and high-speed Internet access.

The most appropriate practice for reaching agreement on complex technical and pricing issues for local loop access is commercial negotiation between the parties involved. However, as experience has demonstrated multiple cases where regulatory intervention is necessary due to imbalance in the negotiation power between the new entrant and those market players having SMP, and due to the lack of other possible alternatives, it should be expected that the role of national regulatory authorities (NRAs) will be crucial for the future (European Parliament & European Council, 2002b). NRAs may intervene at their own initiatives to specify issues, including pricing, designed to ensure interoperability of services, maximise economic efficiency, and benefit end users. Moreover, cost and price rules for local loops and associated facilities (such as collocation and leased transmission capacity; Eutelis Consult GmbH, 1998) should be cost-oriented, transparent, non-discriminatory, and objective to ensure fairness and no distortion of competition.

CURRENT MEANS OF ACCESS & TECHNICAL IMPLEMENTATIONS: THE WAY FORWARD

It is recommended that NRAs ensure that an operator having "SMP" provides its competitors with the same facilities as those that it provides to itself (or to its associated companies), and with the same conditions and time scales. This applies in particular to the roll-out of new services in the local access network, availability of collocation space, provision of leased transmission capacity for access to collocation sites, ordering, provisioning, quality, and maintenance procedures. However, LLU implies that multiple technical, legal, and economical problems have to be solved simultaneously, and decisions have to be made on all relevant topics, especially when market players cannot find commonly accepted solutions (European Parliament & European Council, 2000). Physical access should be normally provided to any feasible termination point of the copper local loop where the

new operator can collocate and connect its own network equipment and facilities to deliver services to its customers.

Theoretically, collocating companies should be allowed to place any equipment necessary to access (European Parliament & European Council, 2002a) the unbundled local loop using available collocation space, and to deploy or rent transmission links from there up to the point of the presence of the new entrant. Furthermore, they should be able to specify the types of collocation available (e.g., shared, caged or cageless, physical or virtual) and to provide information about the availability of power and air-conditioning facilities at these sites with rules for the subleasing of collocation space. NRAs will supervise the entire process to guarantee full appliance of the EU law requirements.

According to the technical approaches proposed (Squire, Sanders, & Dempsey L.L.P., 2002), three ways of access to the local loop of twisted copper pairs can be considered (European Commission, 2000a, 2000b). These can be evaluated (and applied) under certain well-defined criteria based either on technical feasibility or the need to maintain network integrity (OECD, 2003). These distinct solutions can provide complementary means of access and solve various operational aspects in terms of time to market, subscriber take rate, the availability of a second subscriber line, local exchange-node size, spectral compatibility between systems (due to cross-talk between copper pairs), and, the availability of collocation space and capacity in the exchange (Federal Communications Commission, 2001). The different means of access are listed as follows.

Full Unbundling of the Local Loop

In this case, the copper pair is rented to a third party for its beneficiary and exclusive use under a bilateral agreement with the incumbent. The new entrant obtains full control of the relationship with its end user for the provision of full-range telecommu-

Figure 1(a). A simple case of full local loop unbundling

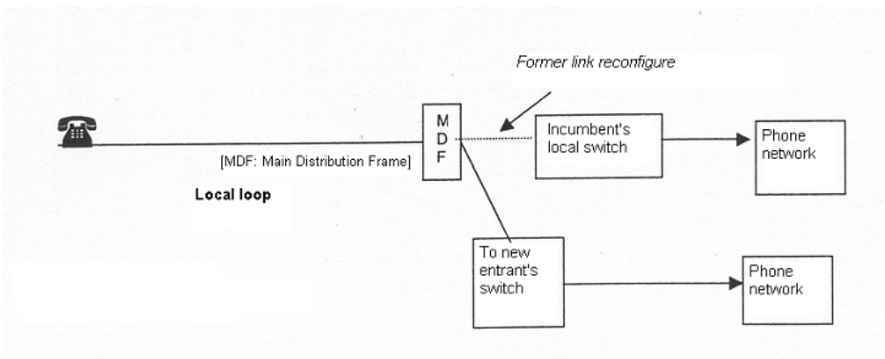
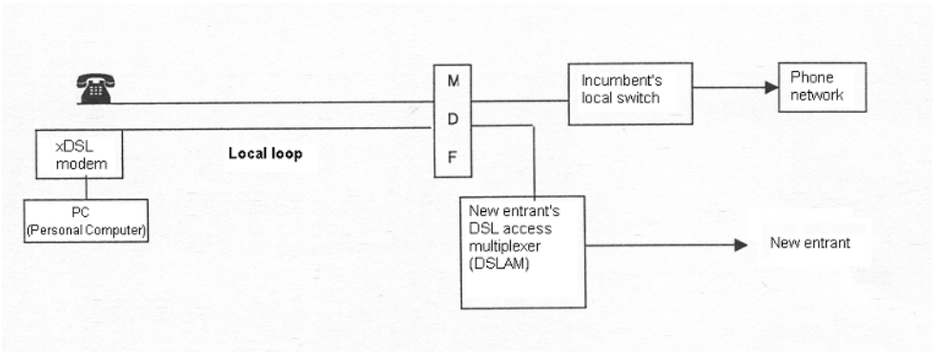


Figure 1(b). A case of full LLU via the use of an xDSL modem



nication services over the local loop, including the deployment of digital subscriber-line systems for high-speed data applications. This option gives the new entrant exclusive use of the full frequency spectrum available on the copper line, thus enabling the most innovative and advanced DSL technologies and services, that is, data rates of up to 60 Mbit/s to the user using VDSL (very high speed DSL). Work on standardizing VDSL is currently taking place in the International Telecommunications Union (ITU) and the European Telecommunications Standards Institute (ETSI).

Figure 1a provides an example where the customer wishes to change telephone and/or leased-line service providers, and the new entrant benefits from "full" unbundling to provide competitive services (probably including multiservice voice and data offerings).

Figure 1b is an alternative case where the new entrant uses full LLU to provide high-speed data service to a customer over a second line using any type of xDSL modems. (In this case, the customer retains the incumbent as the provider of telephone services in the first line.)

Shared Use of the Copper Line

In this case, the incumbent operator continues to provide telephone service using the lower frequency part of the spectrum, while the new entrant delivers high-speed data services over the same copper line using its own high-speed asymmetric-DSL (ADSL) modems. Telephone traffic and data traffic are separated through a splitter before the incumbent's switch. The local loop remains connected to, and part of, the public switched telephone network (PSTN).

The ITU has worked out technical specifications for ADSL full rate—with speeds up to 8 Mbit/s downstream and 1 Mbit/s upstream—in a relevant recommendation (ITU-T, 1997a). This includes a number of country-specific variants in order to accommodate regional local loop infrastructure differences. ADSL can achieve its highest speeds at a distance of 4 km or less. The connection also allows the provision of voice phone service on the basic frequency band of the same line. In addition, the ITU has elaborated a variant ADSL solution in its G.Lite recommendation (ITU-T, 1997b) that is very easy to deploy in the customer premises because it is splitterless (it needs a very simple serial filter that separates voice from data and does not demand any rewiring at the customer premises). Speeds are up to 1.5 Mbit/s downstream to the user, and 385 kbit/s upstream. Some PC suppliers are already marketing relevant equipment with integrated G.Lite-ADSL modems so that standard universal solutions can be rolled out in large scale in the residential market.

This type of access may provide the most cost-effective solution for a user wishing to retain telephone service being provided by the incumbent, but seeking fast Internet service from an Internet service provider (ISP) of his or her choice. The "shared use" provides the feature that different services can be ordered independently from different providers.

Figure 2 provides a relevant example where the new entrant supplies the customer with an ADSL modem for connection, and installs a DSL access multiplexer (which combines ADSL modems and a network interface module) on the incumbent's premises based on a collocation agreement. (The interface between the incumbent's system and the new entrant is at Point C.)

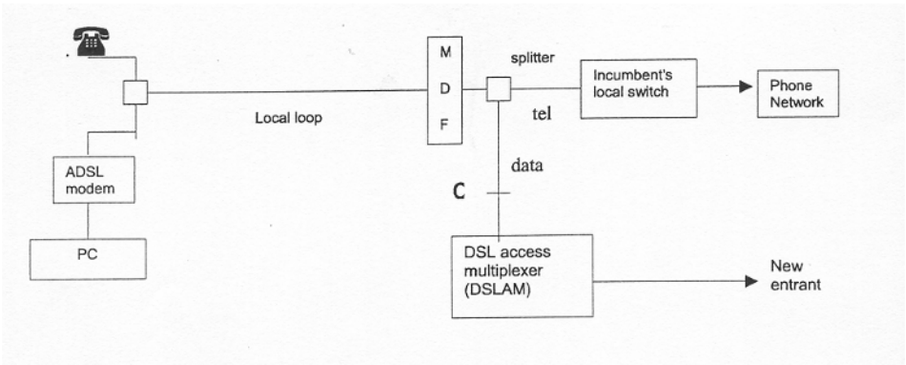
High-Speed Bit-Stream Access or Service Unbundling

This case refers to the situation where the incumbent installs a high-speed access link to the customer premises (e.g., by installing its preferred ADSL equipment and configuration in its local access network) and then makes this access link available to third parties, enabling them to provide high-speed services to customers (European Telecommunications Platform, 2001).

The incumbent may also provide transmission services to its competitors to carry traffic to a higher level in the network hierarchy where new entrants may already have a point of presence (e.g., a transit switch location). Thus, alternative operators can provide services to their end users on a circuit- or switched-service basis. This type of access does not actually entail any unbundling of the copper pair in the local loop (but it may use only the higher frequencies of the copper local loop, as in the case of "shared use").

For a new market player, the problem in exploiting access to unbundled copper pairs is that it entails building out its core network to the incumbent's local

Figure 2. Case of shared use

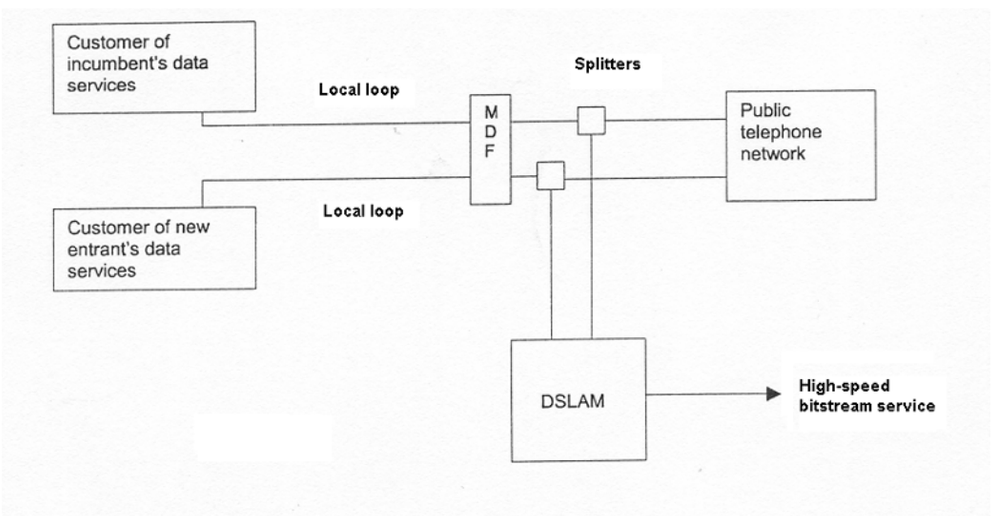


exchanges where the copper pairs are terminated; however, this option, when combined with a transmission service that delivers traffic to the new entrant's point of presence, can be attractive, particularly in the early stage of the newcomer's network deployment. In addition, the "bit-stream access" can be also attractive for the incumbent operator in that it does not involve physical access to copper pairs and so allows for a higher degree of network optimization.

Figure 3 provides an example where two customers continue to receive telephone services from the incumbent operator. The incumbent can dispose a high-speed access link to some third parties. The incumbent may also provide transmission services to its competitors (e.g., by using its ATM [Asynchronous Transfer Mode] or IP [Internet Protocol] network) to carry competitors' traffic from the DSLAM to a higher level in the network hierarchy.

As for the potential application of these three distinct ways of access to the local loop, the European Commission has considered all of them as "complementary"; that is, they should be evaluated in parallel to strengthen competition and improve choice for all users by allowing the market to decide (OFCOM, 2004) which offering best meets users needs, taking into account the evolving demands of users and the technical and investment requirements for market players (OFCOM, 2003). However, the obligation to provide unbundled access to the local

Figure 3. Case of high-speed bit-stream access



loop does not imply that "SMP" operators have to install entirely new local network infrastructure specifically to meet beneficiaries' requests (European Parliament & European Council, 2000).

The development of technical specifications to implement LLU is very complex. Conditions for the unbundled access to the local loop, independent of the particular method used, may contain various distinct technical information (European Commission, 2000a, 2000b; Eutelis Consult GmbH, 1998; Ödling et al., 2000). First of all, it should be absolutely necessary to specify the network elements to which access is offered. This option may include the following: (i) access to raw-copper local loops (copper terminating at the local switch) and subloops (copper terminating at the remote concentrator or equivalent facility) in the case of "full unbundling", (ii) access to nonvoice frequencies of a local loop in the case of "shared access" to the local loop, and (iii) access to space within a main distribution frame site of the notified operator for the attachment of DSLAMs, routers, ATM multiplexers, remote switching modules, and similar types of equipment to the local loop of the incumbent operator.

Another significant perspective refers to the possibility for "availability" and takes into account all relevant details regarding local network architecture, information concerning the locations of physical access sites, and the availability of copper pairs in specific parts of the access network.

The successful provision of LLU will also implicate the explicit definition of various technical conditions, such as technical characteristics of copper pairs in the local loop, lengths, wire diameters, loading coils and bridged taps of the copper infrastructure, and line-testing and conditioning procedures. Other relevant information will include specifications for DSL equipment, splitters, and so forth (with reference to existing international standards or recommendations), as well as usage restrictions, probable spectrum limitations, and electromagnetic compatibility (EMC) requirements designed to prevent interference with other systems.

CONCLUSION

In the framework for the promotion of an advanced, harmonised, and competitive European electronic communications market, offering users a wide choice for a full range of communications services (also including broadband multimedia and local-access high-speed Internet services), local loop unbundling can be a necessary pre-condition (OVUM, 2003) for the healthy development of the relevant market(s) (Chochliouros & Spiliopoulou-Chochliourou, 2002). In particular, recent European regulatory measures have supported the perspective of unbundled access to the copper local loop of fixed operators having significant market power under transparent, fair, and non-discriminatory conditions. Significant progress has been achieved up to the present day, although various problems still exist mainly due to the great complexity of the relevant technical issues (Frantz, 2002). To supersede this obstacle, three alternative LLU methods are currently offered, each one with distinct advantages and different choices for both operators and users or consumers.

The European Commission has evaluated LLU as a means to encourage long-term infrastructure competition (European Commission, 2003) by allowing entrants to "test out" the market before building their own infrastructure and, consequently, to develop infrastructures that promote the growth of electronic communications and e-commerce innovations directly to end users. Thus, the corresponding sectors may offer multiple business opportunities to all market players involved.

Local loop unbundling will complement the recent provisions in EU law, especially to guarantee universal service and affordable access for all citizens by enhancing competition, ensuring economic efficiency, and bringing maximum benefit to users in a secure, harmonised, and timely manner.

REFERENCES

Chochliouros, I., & Spiliopoulou-Chochliourou, A. (2002). Local loop unbundling policy measures as an initiative factor for the competitive development of the European electronic communications markets. *The Journal of the Communications Network: TCN, 1*(2), 85-91.

Chochliouros, I., & Spiliopoulou-Chochliourou, A. (2003). Innovative horizons for Europe: The new European telecom framework for the development of

modern electronic networks & services. *The Journal of the Communications Network: TCN, 2*(4), 53-62.

European Commission. (1999). *Communication on the 1999 communications review: Towards a new framework for electronic communications [COM (1999) 539, 10.11.1999].* Brussels, Belgium: European Commission.

European Commission. (2000a). *Communication on unbundled access to the local loop: Enabling the competitive provision of a full range of electronic communication services, including broadband multimedia and high-speed Internet [COM (2000) 394, 26.07.2000].* Brussels, Belgium: European Commission.

European Commission. (2000b). *Recommendation 2000/417/EC on unbundled access to the local loop: Enabling the competitive provision of a full range of electronic communication services including broadband multimedia and high-speed Internet [OJ L156, 29.06.2000, 44-50].* Brussels, Belgium: European Commission.

European Commission. (2001). *Communication on the seventh report on the implementation of the telecommunications regulatory package [COM (2001) 706, 26.11.2001].* Brussels, Belgium: European Commission.

European Commission. (2003). *Communication on the ninth report on the implementation of the telecommunications regulatory package [COM (2003) 715, 19.11.2003].* Brussels, Belgium: European Commission.

European Parliament & European Council. (1997). *Directive 97/33/EC on interconnection in telecommunications with regard to ensuring universal service and interoperability through application of the principles of open network provision (ONP) [OJ L199, 26.07.1997, 32-52].* Brussels, Belgium: European Commission.

European Parliament & European Council. (2000). *Regulation (EC) 2887/2000 on unbundled access to the local loop [OJ L336, 30.12.2002, 4-8].* Brussels, Belgium: European Commission.

European Parliament & European Council. (2002a). *Directive 2002/19/EC on access to, and interconnection of, electronic communications networks and associated facilities (Access directive) [OJ L108, 24.04.2002, 7-20].* Brussels, Belgium: European Commission.

European Parliament & European Council. (2002b). *Directive 2002/21/EC on a common regulatory framework for electronic communications networks and services (Framework directive) [OJ L108, 24.04.2002, 33-50].* Brussels, Belgium: European Commission.

European Telecommunications Platform (ETP). (2001). *ETP recommendations on high-speed bitstream services in the local loop.* Brussels, Belgium: European Telecommunications Platform.

Eutelis Consult GmbH. (1998). *Recommended practices for collocation and other facilities sharing for telecommunications infrastructure* (Study for DG XIII of the European Commission, Final report). Brussels, Belgium: European Commission.

Federal Communications Commission (FCC). (2001). *In the matter of review of the section 251 unbundling obligations of incumbent local exchange carriers* (CC Docket No. 01-338). Washington, DC: Federal Communications Commission.

Frantz, J. P. (2002). The failed path of broadband unbundling. *The Journal of the Communications Network: TCN, 1*(2), 92-97.

ITU-T. (1997a). *Recommendation G.992.1: Asymmetric digital subscriber line (ADSL) transceivers*. Geneva, Switzerland: International Telecommunications Union (ITU).

ITU-T. (1997b). *Recommendation G.992.2: Splitterless asymmetric digital subscriber line (ADSL) transceiver*. Geneva, Switzerland: International Telecommunications Union (ITU).

Ödling, P., Mayr, B., & Palm, S. (2000, May). The technical impact of the unbundling process and regulatory action. *IEEE Communications Magazine, 38*(5), 74-80.

OECD. (2003). Working party on telecommunications and information services policies. In *Developments in local loop unbundling (DSTI/ICCP/TISP(2002)5/FINAL, JT00148819).* Paris, France: Organisation for Economic Co-operation and Development (OECD).

OFCOM. (2003). *Local loop unbundling fact sheet.* London, United Kingdom: OFCOM.

OFCOM. (2004). *Review of the wholesale local access markets.* London, United Kingdom: OFCOM.

OVUM. (2003). *Barriers to competition in the supply of electronic communications networks and services: A final report to the European Commission.* Brussels, Belgium: European Commission.

Squire, Sanders, & Dempsey L.L.P. (2002). *Legal study on part II of local loop unbundling sectoral inquiry* (Contract No. Comp. IV/37.640). Brussels, Belgium: European Commission.

KEY TERMS

Asymmetric DSL (ADSL): A DSL technology that allows the use of a copper line to send a large quantity of data from the network to the end user (downstream data rates up to 8 Mbit/s), and a small quantity of data from the end user to the network (upstream data rates up to 1 Mbit/s). It can be used for fast Internet applications and video-on-demand.

Bandwidth: The physical characteristic of a telecommunications system indicating the speed at which information can be transferred. In analogue systems it is measured in cycles per second (Hertz), and in digital systems it is measured in binary bits per second (bit/s).

Broadband: A service or connection allowing a considerable amount of information to be conveyed, such as video. It is generally defined as a bandwidth of over 2 Mbit/s.

Copper Line: The main transmission medium used in telephony networks to connect a telephone or other apparatus to the local exchange. Copper lines have relatively narrow bandwidth and limited ability to carry broadband services unless combined with an enabling technology such as ADSL.

DSL (Digital Subscriber Loop): The global term for a family of technologies that transform the copper local loop into a broadband line capable of delivering multiple video channels into the home. There are a variety of DSL technologies known as xDSL; each type has a unique set of characteristics in terms of performance (maximum broadband capacity), distance over maximum performance (measured from the switch), frequency of transmission, and cost.

Local Loop: The access network connection between the customers' premises and the local public switched telephony network (PSTN) exchange, usually a loop comprised of two copper wires. In fact, it is the physical twisted metallic pair circuit connecting the network termination point at the subscriber's premises to the main distribution frame or equivalent facility in the fixed public telephone network.

Main Distribution Frame (MDF): The apparatus in the local concentrator (exchange) building where the copper cables terminate and where cross-connection to other apparatuses can be made by flexible jumpers.

Public Switched Telephony Network (PSTN): The complete network of interconnections between telephone subscribers.

Very High Speed DSL (VDSL): An asymmetric DSL technology that provides downstream data rates within the range 13 to 52 Mbit/s, and upstream data rates within the range 1.5 to 2.3 Mbit/s. VDSL can be used for high capacity leased lines as well as for broadband services.

Making Money with Open-Source Business Initiatives

Paul Benjamin Lowry
Brigham Young University, USA

Akshay Grover
Brigham Young University, USA

Chris Madsen
Brigham Young University, USA

Jeff Larkin
Brigham Young University, USA

William Robins
Brigham Young University, USA

INTRODUCTION

Open-source software (OSS) is software that can be used freely in the public domain but is often copyrighted by the original authors under an open-source license such as the GNU General Public License (GPL). Given its free nature, one might believe that OSS is inherently inferior to proprietary software, yet this often is not the case. Many OSS applications are superior or on par with their proprietary competitors (e.g., MySQL, Apache Server, Linux, and Star Office). OSS is a potentially disruptive technology (Christensen, 1997) because it is often cheaper, more reliable, simpler, and more convenient than proprietary software.

Because OSS can be of high quality and capable of performing mission-critical tasks, it is becoming common in industry; the majority of Web sites, for example, use Apache as the Web server. The deployment of OSS is proving to be a productive way to counter the licensing fees charged by proprietary software companies. An organized approach to distributing cost-effective OSS products is intensifying as companies such as RedHat and IBM co-brand OSS products to establish market presence.

From a business perspective, the entire OSS movement has been strategically anti-intuitive because it is based on software developers freely sharing source code—an act that flies in the face of traditional proprietary models. This movement raises two questions this article aims to address: (1) why would individuals write software and share it freely? and (2) how can software firms make money from OSS? Before fully addressing these questions, this article examines the historical development of OSS.

OSS HISTORY

A strategic irony of the software industry is that its foundation rests primarily on OSS principles. Software development in the 1960s and 1970s was steered primarily by government and academia. Software developers working in the field at the time considered it a normal part of their research culture to exchange, modify, and build on one another's software (Von Krogh, 2003). Richard Stallman, a professor and programmer at MIT, was a strong advocate and contributor to this culture of open, collaborative software development. Despite Professor Stallman's influence, MIT eventually stopped exchanging sourcecode with other universities to increase its research funding through proprietary software licensing. Offended by MIT's decision to limit code sharing, Professor Stallman founded the Free Software Foundation in 1985 and developed the General Public

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

License (GPL) to preserve free code sharing (Bretthauer, 2002).

In the formative years of the software industry, Stallman's free software movement grew slowly; in the early 1990s, however, the concept of code sharing grew more rapidly for a couple of reasons. First, "free software" was renamed "OSS," a name that spread rapidly throughout the code-sharing community (Fitzgerald & Feller, 2001). Second, the OSS movement received a boost from the advent of the World Wide Web (WWW). The Web provided an opportunity for Internet users to quickly and conveniently share their code.

WHY DEVELOPERS WRITE OSS

The majority of OSS software developers fall into one of the following three categories: freelancers, software enthusiasts, or professionals. Freelancers enjoy the challenges associated with developing OSS and providing services to the OSS community to further their own careers. When freelancers create modules of code, they often include their contact information inside the modules (Lerner & Tirole, 2002). This allows businesses to contact the developers to request their future services.

Software enthusiasts are people who contribute to OSS simply out of the joy and challenge of doing so, with little regard for professional advancement. Enthusiasts are often university students who want to participate in the development of free software and who receive personal gratification from participating in real-world OSS development projects and gaining the respect of the OSS community.

Even though OSS is "free" software, many companies hire professional developers to work on improving OSS code. RedHat, a Linux support company, hires developers to fix bugs in OSS code and to create new applications (Lerner & Tirole, 2002). Other companies hire OSS developers because their systems run OSS applications and they need developers to customize the code for specific business purposes. Table 1 summarizes the different motivations for joining OSS projects and shows them on a spectrum of intrinsic and extrinsic motivations.

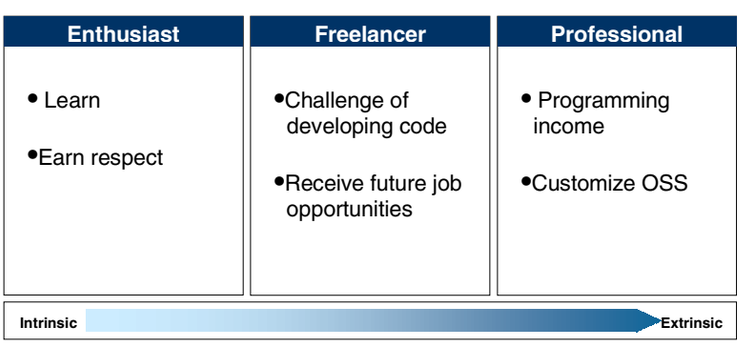
SOFTWARE DEVELOPMENT ECONOMICS

Proprietary software

The strategic motivation behind the creation of proprietary software is to set up high switching costs for consumers. For such companies their developers' resulting source code becomes the company's intellectual property and an unshared key company asset. Once customers purchase proprietary software, they must pay for updates continually to keep the software current, and often to receive full customer support (Delong & Froomkin, 2000). Most customers will pay these fees because of the lock in that occurs from the often costly prohibitive tradeoff of implementing a completely new system.

Microsoft is an example of a company that has succeeded in proprietary software, largely because they have a focused strategy of selling complementary products and services to their installed base of Windows users (Shapiro & Varian, 1998): Offering

Table 1. Developer motivations



complementary goods that run on Windows (e.g., Office) increases profitability and successfully enhances the buyer relationship while encouraging customer entrenchment.

Proprietary software development is rigidly structured. Development begins with an end product in mind, and the new product often integrates with other products the company is currently selling. Project leaders create development plans, set deadlines, and coordinate teams to develop modules of the new software product. Successful proprietary software companies are also able to develop new technologies in exceptionally short time frames and to place their products in the market faster than their competitors. Products that meet the strict demands of end users succeed and increase customer satisfaction.

The downside of proprietary software development is that it comes at a tremendous internal cost (Lederer & Prasad, 1993); meanwhile, the industry is experiencing increasing pressures to decrease costs. Companies must invest heavily in research and development (R&D), human capital, information technology, marketing, brand development, and physical manufacturing of the products. They must continually innovate and develop updated versions of existing products, or create entirely new products. To compensate for these costs, proprietary software companies have high-priced products. Some software costs are so high that many businesses question whether the software is worth it.

OSS

The economics of OSS differ significantly in that OSS is developed in a loose marketplace structure. The development process begins when a developer presents an idea or identifies a need for an application with specific functionality (Johnson, 2002). OSS software development typically has a central person or body that selects a subset of developed code for an "official" release and makes it widely available for distribution. OSS is built by potentially large numbers of volunteers in combination with for-profit participants (Von Krogh, 2003). Often no system-level design or even detailed design exists. Developers work in arbitrary locations, rarely or never meet face to face, and often coordinate their activity through e-mail and bulletin boards. As participants make changes to the original application, the central person or body leading the development selects code changes, incorporates them into the application, and officially releases the next version of the application. Table 2 compares OSS to proprietary development.

OSS BUSINESS MODELS

A business model is a method whereby a firm builds and uses resources to provide a value-added proposition to potential customers (Afuah & Tucci, 2000).

Table 2. OSS development vs. proprietary development

OSS	Proprietary Software
Similarities	
Building brand name and reputation increases software use	
Revenue is generated from supporting software, creating new applications for software, and certifying software users	
Differences	
Code developed outside of company for free	Developers are paid to program code
Source code is open for public use.	Source code is kept in company
People use program without paying any license fees.	Users pay license fees to use the software
Updates are free and users are allowed flexibility in using them	People are locked in using specific software and have to pay for updates
Code is developed for little internal cost	Code is costly to create internally

OSS business models are based on providing varied services that cater to cost-sensitive market segments and provide value to the end user by keeping the total cost of ownership as low as possible (Hecker, 1999). OSS-based companies must provide value-added services that are in demand, and they must provide these services at cost-sensitive levels. OSS is a strategic threat to proprietary software, because one of the most effective ways to compete in lock-in markets is to "change the game" by expanding the set of complementary products beyond those offered by rivals (Shapiro & Varian, 1998). OSS proponents are trying to "change the game" with new applications of the following business models (Castelluccio, 2000): support sellers, loss leaders, code developers, accessorizers, certifiers, and tracking service providers.

Support Sellers

Support sellers provide OSS to customers for free, except for a nominal packaging and shipping fee, and instead charge for training and consulting services. They also maintain the distribution channel and branding of a given OSS package. They provide value by helping corporations and individuals install, use, and maintain OSS applications. An example of a support seller is RedHat, which provides reliable Linux solutions.

To offer such services, support sellers must anticipate and provide services that will meet the needs of businesses using OSS. To offer reliable and useful consulting services, support sellers must invest heavily in understanding the currently available OSS packages and developing models to predict how these OSS applications will evolve in the future (Krishnamurthy, 2003).

This model has strengths in meeting the needs for outsourcing required IT services, which is the current market trend (Lung Hui & Yan Tam, 2002). OSS provides companies an opportunity to reduce licensing costs by allowing companies to outsource the required IT support to support sellers. Likewise, the marketplace structure of OSS development adds significant uncertainty to the future of OSS applications. Risk-adverse companies often do not want to invest in specialized human capital, and support sellers help mitigate these risks.

One drawback of this model is that consulting companies often fall prey to economic downturns, during which potential clients reduce outsourcing to consultants. This cycle is compounded for the software industry, since a poor economy results in cost cutting and an eventual reduction in IT spending.

Loss leaders

Loss-leader companies write and license proprietary software that can run on OSS platforms (Castelluccio, 2000). An example of a loss leader is Netscape, which gives away its basic Web-browser software but then provides proprietary software or hardware to ensure compatibility and allow expanded functionality. The loss-leader business model adds value by providing applications to companies that have partially integrated OSS with their systems (Hecker, 1999). Companies often need specific business applications that are unavailable in the OSS community, or they desire proprietary applications but wish to avoid high platform-licensing costs.

To leverage the integration of OSS with proprietary software, loss leaders need to assemble a team of highly skilled developers, create an IT infrastructure, and develop licensable applications. The major costs of this business model arise from payroll expenses for a development team, R&D costs, marketing, and, to a lesser extent, patenting and manufacturing.

This model's strength is that it provides a solution for the lack of business applications circulating in the OSS community. The loss leader model fills the gap between simpler available OSS applications, such as word processors, and more complex applications that are unavailable in the OSS community.

A weakness of this model is the risk of disintermediation. As time passes and OSS coding continues to grow and expand, more robust and complex applications will be developed. However, the developers of these applications will have to cope with the speed and efficiency of proprietary software development.

Code Developers

The code development model addresses some of the limitations of the loss-leader model. Code development companies generate service revenue through

on-demand development of OSS. If a firm cannot find an OSS package that meets its needs for an inventory management system, for example, the firm could contract with a code development company to the basic application (Johnson, 2002). The code development company could then distribute this application to the OSS community and act as the development project's leader. The code development company would track the changes made to the basic source code by the OSS community and integrate those changes into its product. The company would periodically send its customers product updates based on changes accepted from the OSS community.

The necessary assets and associated costs required by this model are similar to those in the proprietary software model, including a team of programmers, IT infrastructure, and marketing. However, the code developer needs to develop only a basic application. Once the basic software is developed, the OSS community provides further add-ons and new features (Johnson, 2002), which decrease the R&D costs for the company acting as project leader. Yet the code development team needs to have the necessary IT infrastructure to lead the OSS community in the application's evolution, incorporate new code, and resubmit new versions to its customers.

This model's strength is its longevity. The code development model overcomes the risk of disintermediation by basing its revenue generation on initiating OSS applications and maintaining leadership over their evolution; it does not focus on privatizing the development and licensing of applications.

This model's weakness is the risk of creating an application of limited interest to the OSS community. A possible solution to this problem would be an offer from the company leading the development process to reward freelance developers for exceptional additions to the application's original code.

Accessorizers

Accessorizers companies add value by selling products related to OSS. Accessorizers provide a variety of different value-added services, from installing Linux OS on their clients' hardware to writing manuals and tutorials (Hecker, 1999; Krishnamurthy, 2003). For example, O'Reilly & Associates, Inc. writes manuals for OSS and produces downloadable copies of Perl, a programming language.

One strength of this model is that it provides the new manuals and tutorials that the constantly changing nature of the OSS market requires. Another strength is its self-perpetuating nature: as more manuals and tutorials are produced, more people will write and use OSS applications, increasing the need for more manuals and tutorials.

This models' weakness is the difficulty of staying current with the many trends with the OSS community. This difficulty creates the risk of investing in the wrong products or producing too much inventory that is quickly outdated.

Certifiers

Certifiers establish methods to train and certify students or professionals in an application. Certificate companies like CompTIA generate revenue through training programs, course materials, examination fees, and certification fees. These programs provide value to the individuals enrolled in the certification programs and businesses looking for specific skills (Krishnamurthy, 2003). Certification helps the OSS industry by creating benchmarks, expectations, and standards employers can use to evaluate and hire employees based on specific skill sets.

Certification has long-term profit potential since most certification programs require recertification every few years due to continuing education requirements. Businesses value certification programs because they are a cost-effective way to train employees on new technologies. Certifiers, who achieve first-mover advantage, become trendsetters for the entire industry, increasing barriers to entry into the certification arena.

One downside of this model is the significant startup costs. Certifiers need to find qualified individuals to create manuals, teach seminars, and write tests. Certifiers must also survey businesses to discern which parts of specific applications are most important, and which areas need the greatest focus during training. Certifiers also need to gain substantial credibility through marketing and critical mass or their tests have little value. Increasing company name recognition and building a reputation in the certification arena can be an expensive and long process.

This model also faces the threat of disintermediation. Historically, certification programs have evolved into not-for-profit organizations, such

as the AICPA in accounting, or the ISO 9000 certification in operations. The threat of obsolescence is another major weakness. In the 1970s, FORTRAN or COBOL certification may have been important (Castelluccio, 2000), but they have since become obsolete. Certifiers specializing in certain applications must be constantly aware of the OSS innovation frontier and adjust their certification options appropriately.

Tracking Service Providers

The tracking-services business model generates revenue through the sale of services dedicated to tracking and updating OSS applications. For example, many companies have embraced Linux to cut costs; however, many of these same companies have found it difficult to maintain and upgrade Linux because of their lack of knowledge and resources. Tracking-services companies, like Sourceforge.net and FreshMeat.net, sell services to track recent additions, define source code alternatives, and facilitate easy transition of code to their customers' systems.

A strength of this model is its ability to keep costs low by automating the majority of the work involved in tracking while still charging substantial subscription and download fees. However, these services must have Web-based interfaces with user-friendly download options, and they also must develop human and technological capabilities that find recent updates and distinguish between available alternatives.

A weakness of this model is low barriers to entry. This information-services model can be replicated with a simple Web interface and by spending time on OSS discussion boards and postings, creating the possibility of such services becoming commoditized. Table 3 summarizes some of the differences between the OSS business models.

CONCLUSION

The market battle between OSS and proprietary software has just begun. This battle could be termed a battle of complementary goods and pricing. For example, the strategies between Microsoft and RedHat are similar in that they both need a large, established user base that is locked in and has access to a large array of complementary goods and services. The key differences in their strategies are in their software development process, software distribution, intellectual property ownership, and pricing of core products and software. It will be increasingly important for OSS companies to track the competitive response of proprietary companies in combating the increasing presence of OSS.

Moreover, the OSS movement has begun to make inroads into the governments in China, Brazil, Australia, India, and Europe. As whole governments adopt OSS the balance of power can shift away from proprietary providers. This also provides the opportunity to develop a sustainable business model that caters only to the government sector. Similarly, for-

Table 3. OSS models

Business Model	Assets	Costs	Revenue Model
Support Sellers	Human capital, supporting infrastructure, contracts	Payroll, IT, marketing and brand development	Training, consulting
Loss Leaders	Human capital, supporting infrastructure, software	Payroll, IT cost, marketing and brand development, R&D, software manufacturing	Licenses
Accessorizers	Human capital, supporting infrastructure	Payroll, printing material machines, training, software	Book Sales
Code developers	Human capital, software - technology tracking, database	Payroll, IT, marketing (Corporations), marketing (Freelancers)	Corporations that pay for service
Certifiers	Human capital, IT, Certification program	Certification program development, payroll	Tests, certificates
Tracking-service providers	Human capital, Software-technology tracking, Databases	Payroll, IT, marketing (Corporations)	Corporations that pay for service

mulating business models for corporations and educational institutions may be another fruitful opportunity.

The recent government regulations associated with the Sarbanes-Oxley Act and other financial-reporting legislation are important trends. These regulations require significant research in the area of internal control reporting on OSS applications. It is likely the collaborative and less proprietary nature of OSS could help with this reporting. If this reporting can be done with more assurance than provided by proprietary applications, OSS providers can gain further advantage.

REFERENCES

Afuah, A. & Tucci, C. (2000). *Internet business models and strategies: Text and cases.* McGraw-Hill Higher Education.

Bretthauer, D. (2002). Open source software: A history. *Information Technology & Libraries, 21*(1), 3-10.

Castelluccio, M. (2000). Can the enterprise run on free software? *Strategic Finance, 81*(9), 50-55.

Christensen, C.M. (1997). *The innovator's dilemma: When new technologies cause great firms to fail.* Harvard Business School Press.

Delong, J.B. & Froomkin, A.M. (2000). Beating Microsoft at its own game. *Harvard Business Review, 78*(1), 159-164.

Fitzgerald, B. & Feller, J. (2001). Guest editorial on open source software: Investigating the software engineering, psychosocial and economic issues. *Information Systems Journal, 11*(4), 273-276.

Hecker, F. (1999). Setting up shop: The business of open-source software. *IEEE Software, 16*(1), 45-51.

Johnson, J.P. (2002). Open source software: Private provision of a public good. *Journal of Economics & Management Strategy, 11*(4), 637-662.

Krishnamurthy, S. (2003). A managerial overview of open source software. *Business Horizons, 46*(5), 47-56.

Lederer, A.L. & Prasad, J. (1993). Information systems software cost estimating: A current assessment. *Journal of Information Technology, 8*(1), 22-33.

Lerner, J. & Tirole, J. (2002). Some simple economics of open source. *Journal of Industrial Economics, 50*(2), 197-234.

Lung Hui, K. & Yan Tam, K. (2002). Software functionality: A game theoretic analysis. *Journal of Management Information Systems (JMIS), 19*(1), 151-184.

MacCormack, A. (2001). Product-development practices that work: How Internet companies build software. *MIT Sloan Management Review, 42*(2), 75-84.

Shapiro, C. & Varian, H.R. (1998). *Information rules: A strategic guide to the network economy.* Harvard Business School Press.

Von Krogh, G. (2003). Open-source software development. *MIT Sloan Management Review, 44*(3), 14-18.

KEY TERMS

Copyright: A legal term describing rights given to creators for their literary and artistic works. See World Intellectual Property Organization at www.wipo.int/about-ip/en/copyright.html.

General Public License (GPL): License designed so that people can freely (or for a charge) distribute copies of free software, receive the source code, change the source code, and use portions of the source code to create new free programs.

GNU: GNU is a recursive acronym for "GNU's Not Unix." The GNU Project was launched in 1984 to develop a free Unix-like operating system. See www.gnu.org/.

Open-source Software (OSS): Software that can be freely used in the public domain, but is often copyrighted by the original authors under an open-source license such as the GNU GPL. See the Open Source Initiative at www.opensource.org/docs/definition_plain.php.

Malware and Antivirus Procedures

Xin Luo
Mississippi State University, USA

Merrill Warkentin
Mississippi State University, USA

INTRODUCTION

The last decade has witnessed the dramatic emergence of the Internet as a force of inter-organizational and inter-personal change. The Internet and its component technologies, which continue to experience growing global adoption, have become essential facilitators and drivers in retailing, supply chain management, government, entertainment, and other processes. However, this nearly-ubiquitous, highly-interconnected environment has also enabled the widespread, rapid spread of malware, including viruses, worms, Trojan horses, and other malicious code. Malware is becoming more sophisticated and extensive, infecting not only our wired computers and networks, but also our emerging wireless networks. Parallel to the rise in malware, organizations have developed a variety of antivirus technologies and procedures, which are faced with more challenging tasks to effectively detect and repair current and forthcoming malware. This article surveys the virus and antivirus arena, discusses the trends of virus attacks, and provides solutions to existing and future virus problems (from both technical and managerial perspectives).

BACKGROUND

Global networks and client devices are faced with the constant threat of malware attacks which create burdens in terms of time and financial costs to prevent, detect, repel, and especially to recover from. Viruses represent a serious threat to corporate profitability and performance, and accordingly, this constant threat of viruses and worms has pushed security to the top of the list of key issues (Palmer, 2004). Computer virus attacks cost global businesses an estimated $55 billion in damages in 2003. Companies lost roughly $20 billion to $30 billion in 2002 from the virus attacks, up from about $13 billion in 2001 (Tan, 2004). Sobig.F virus, for instance, caused $29.7 billion in economic damage worldwide (Goldsborough, 2003) (see Tables 1 and 2). Mydoom and its variants have infected 300,000 to 500,000 computers (Salkever, 2004), and Microsoft has offered $250,000 for information leading to the arrest of the worm writer (Stein, 2004). The most recent Sasser worm has infected millions of users including American Express (AMEX), Delta Airline Inc. and some other Fortune 500 companies. And Netsky-D worm has caused $58.5 million in damages worldwide (Gaudin, 2004). Furthermore, CSX, the biggest railroad company in USA, had to suspend its services in the metropolitan Washington DC area due to the activity of Nachi worm in 2003. Air Canada cancelled its flights because its network failed to handle the amount of traffic generated by the Nachi worm. Many organizations, whether high, medium, or low profile, such as The Massachusetts Institute of Technology (MIT) and the U.S. Department of Defense have been the victims of viruses and worms. (For more perspective on the financial impact of malware, see Tables 1-4 and Figures 1-2.)

Since the early 1990s, computer viruses have appeared in the world of IT and become more changeable and destructive in recent years, as today's computers are far faster, and as the vulnerabilities of the globally connected Internet are being exploited. Today's hackers can also easily manipulate existing viruses so that the resulting code might be undetectable by antivirus application; they may even insert malicious code into files with no discernable trace to be found, regardless of digital forensics examiner's competence and equipment (Caloyannides, 2003).

A virus is a piece of programming code usually disguised as something else that causes some unexpected and usually undesirable event. Viruses, unlike

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. Top 10 viruses in 2003

Rank	Virus	Percentage of reports
1	W32/Sobig-F	19.9%
2	W32/Blaster-A	15.1%
3	W32/Nachi-A	8.4%
4	W32/Gibe-F	7.2%
5	W32/Dumaru-A	6.1%
6	W32/Sober-A	5.8%
7	W32/Mimail-A	4.8%
8	W32/Bugbear-B	3.1%
9	W32/Sobig-E	2.9%
10	W32/Klez-H	1.6%
Others		25.1%

Source: Sophos.com (http://www.sophos.com)

Table 3. Annual global financial impact of major virus attacks 1995-2003

Year	Impact ($U.S.)
2003	$13.5 Billion
2002	11.1 Billion
2001	13.2 Billion
2000	17.1 Billion
1999	12.1 Billion
1998	6.1 Billion
1997	3.3 Billion
1996	1.8 Billion
1995	500 Million

Source: Computer Economics

http://www.computereconomics.com/article.cfm?id=936

Table 2. Number of security incidents reported to CERT from 1995-2003

1995	1996	1997	1998	1999	2000	2001	2002	2003
2,412	2,573	2,134	3,374	9,859	21,756	52,658	82,094	137,529

Source: Computer Economics (http://www.computereconomics.com/article.cfm?id=936)

Figure 1. Number of security incidents reported to CERT from 1995-2003

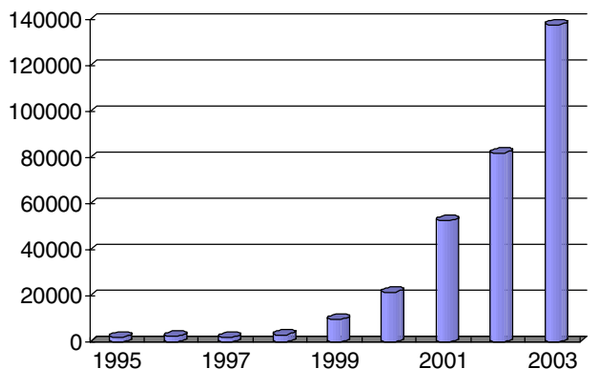
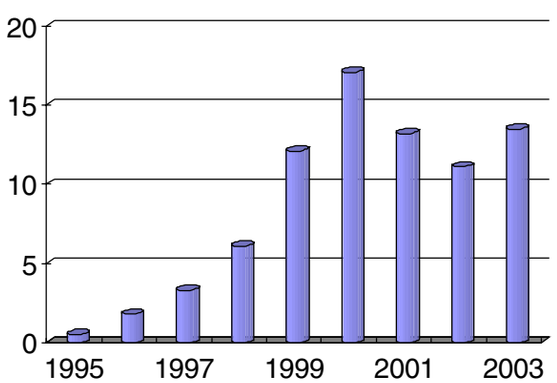


Figure 2. Annual global financial impact of major virus attacks 1995-2003



worms, must attach themselves to another file (typically an executable program file, but can infect dozens of file types, including scripts and data files with embedded macros) in order to propagate. When the host file is executed, the virus's programming is also executed in the background. Viruses are often designed so that they can automatically spread to other computer users (Harris, 2003) by various

Table 4. Global financial impact of major virus attacks since 1999

Year	Code Name	Impact ($U.S.)
2004	MyDoom	$4.0 Billion
2003	SoBig.F	2.5 Billion
2003	Slammer	1.5 Billion
2003	Blaster	750 Million
2003	Nachi	500 Million
2002	Klez	750 Million
2002	BugBear	500 Million
2002	Badtrans	400 Million
2001	CodeRed	2.75 Billion
2001	Nimda	1.5 Billion
2001	SirCam	1.25 Billion
2000	Love Bug	8.75 Billion
1999	Melissa	1.5 Billion
1999	Explorer	1.1 Billion

Source: Computer Economics

http://www.computereconomics.com/article.cfm?id=936

media channels and various methods. Viruses may be transmitted as e-mail attachments, downloaded files, or background scripts, or may be embedded in the boot sector or other files on a diskette or CD. Viruses are usually executed without the computer user's knowledge or choice in the initial stages. (The user may be well aware of them once their latency period ends, and the damage is initiated.) Viruses can be categorized in five types according to various characteristics (Symantec, 2004):

1. **File infector viruses:** infect program files
2. **Boot sector viruses:** infect the system area of a disk
3. **Master boot record viruses:** memory resident viruses that infect disks in the same manner as boot sector viruses
4. **Multi-partite viruses:** infect both boot records and program files
5. **Macro viruses:** infect data files, such as Microsoft Office Suite files with macro script capabilities

Unlike viruses that require the spreading of an infected host file, worms are programs that replicate themselves from system to system without the use of a host file. They can infect various kinds of computer operating systems. For example, the Great Worm, perpetrated by Robert T. Morris, was a program which took advantage of bugs in the Sun Unix sendmail program, Vax programs, and other security loopholes to distribute itself to over 6000 computers on the Internet in its early days. By 2003, the scope and distribution speed of worm attacks had grown to an alarming rate (Arce, 2004).

Trojan Horses are files that are malicious under a disguised cover which often seems desirable for computer users. Unlike viruses, Trojan horses cannot self- replicate. Trojans contain malicious code that can trigger loss or theft of data. Recent Trojan Horses can come in the form of e-mail attachments claiming to be from legitimate source, such as Microsoft security updates, luring people to open the attachments and turning out to disable antivirus and firewall software. The recent explosion of so-called "Phishing" e-mails, disguised as legitimate e-mails (to capture sensitive information for the purposes of identity theft), are a related phenomenon.

MALWARE ATTACK TRENDS

The IT world is experiencing the transition from an old traditional form of viruses and worms to a new and more complicated one. The trend of virus attacks is that fast worms such as the recent Mydoom as well as new blended attacks that combine worms and viruses are the major infective force in the cyber world and will likely become more frequent in years ahead. In general, such viruses are spreading via updated and increasingly sophisticated methods and are capable of damaging more effectively. Since only their creators know how these attacks will launch, IT antivirus teams have encountered extremely difficult predicaments regarding how to proactively prevent the malware disaster and eventually eliminate any malware infection or breach. New virus will broaden its connectivity spectrum, ranging from wired to the newly emerged wireless networks. Particularly, weak 802.11 protocol-based wireless networks are being confronted by increasing attacks.

Between 2001 and 2003, the growth of malware seemingly slowed down. However, on the flip side, new viruses mirror a more constant threat and last longer than in previous years. Notwithstanding the slow down in the growth of new viruses, the prevalence of mass mailing viruses and Internet worms account for the increase in durability. The new viruses are harder to eliminate. The cost of cleaning up after a virus infection has risen. According to ICSA's latest virus prevalence survey, the average cost to companies was $81,000 with $69,000 in the last survey (Roberts, 2003). In 2003, a great number of attacks stem from the combined exploitation of server and workstation vulnerabilities with the characteristics of virus and Trojan horses. By using more efficient attack vectors and, therefore, minimizing the human effort required to deliver attacks and use the compromised systems, the risks related to newly discovered vulnerabilities moved up in the risk measurement scale (Arce, 2004).

Although the new type of viruses are still in infancy, their creators are exploiting the globally connected networks as the hotbed and will combine fast propagation with a destructive payload, such as worms that send private or classified data to an outside location, or destroy or modify data (Rash, 2004).

ANTIVIRUS SOFTWARE

Antivirus software was developed to combat the viruses mentioned above and help computer users have a technologically sanitary environment. Antivirus software is not just about preventing destruction and denial of service (DoS); it's also about preventing hacking and data theft (Carden, 1999). The antivirus software market, which totaled $2.2 billion in 2002, will double to $4.4 billion in 2007 (Camp, 2004). The leading brands of antivirus application brands are Symantec's Norton AntiVirus, McAfee, Trend Micro's PC-cillin, Panda, and F-Secure. Once installed, the software enables auto-protect application that runs constantly in the background of computer, checking incoming and outgoing files and media such as webpages, CDs, diskettes and emails against the virus definitions incorporated into the software to detect any matches. The key to efficient antivirus application is in the updates of virus definitions. Most antivirus software vendors offer updates as often as once a week. The subscription service to enable the updating of virus definitions is typically only about US$20-30 per year.

Today, antivirus applications predominantly operate by means of identifying unique characteristics or certain patterns in the file code that forms a virus. Once identified, this "signature" is distributed, mainly via the software manufacturer's Web site, as the most recent virus definition to people who have licensed and installed antivirus software, allowing the software to update its prior definition in order to recognize, eradicate, or quarantine the malicious code. To monitor e-mails and files moving in and out of a computer as well as Web pages user browses, today's antivirus application typically adopts one or more of the following methods (SolutionsReview.com, 2003):

- **File Scanning:** Scans certain or all files on the computer to detect virus infection. This is the most common scenario that computer users follow to detect and eliminate viruses. Additionally, users can set up schedules to launch automatic virus scanning in the background.
- **E-Mail and Attachment Scanning:** Scans both the scripts of email content and attachments for malware. For example, Norton can detect viruses by analyzing e-mail before passing it to e-mail server for delivery, despite the cumbrance that this would trigger delay of messaging. Today's antivirus applications can even screen malware hidden in compressed packages attached to e-mail messages, such as rar and zip files. Many leading email service providers, such as Yahoo and MSN, are integrating the scanning to e-mail messaging so as to promptly warn users of malicious viruses.
- **Download Scanning:** Simultaneously scans files that are being downloaded from a network. This enables the application to scan not only the directly downloaded files through browser hyperlinks, but also can screen the files via particular downloader.
- **Heuristic Scanning:** Detects virus-like codes/scripts in e-mails and files based on intelligent guessing of typical virus-like code/script patterns and behavior previously analyzed and stored in the application.

- **Active Code Scanning:** Scans active codes like Java, VB script, and ActiveX in Web pages which can be of malicious and do severe damage to the computers. Links in e-mails can invoke active codes in a Web page and do the same damage.

DETECTING MALWARE

Due to their varied nature and constant sophisticated evolution, today's viruses are harder to detect and remove. Firewalls can filter a limited number of damages, but they cannot completely eliminate all types of viruses. Thus, the core activity of detecting malware is the inspection of application behavior. Though effortless, this activity requires constant vigilance. Normally, any unexpected behavior from an application can mirror a sign of a virus or worm at work. For instance, a computer may slow down, stop responding, or crash and restart every few minutes. Sometimes a virus will attack the files we need to start up a computer. All of these are the symptoms that the computer is infected by malware. Also, watching an application with networking monitoring tools, such as Windows Task Manager and Firewall software, can indicate a lot about the traffic going on in the network. When there is anomalous application behavior, such as sudden enormous outward and inward data flow, it is a sign that a virus/worm is propagating.

PREVENTING MALWARE

All security solutions stem from a series of good policies. Good security, especially in the case of worms and viruses, means addressing employee and staff training, physical security, and other cultural changes that allow security technologies to do their best work (Rash, 2004). Upon the establishment of a security policy, we must balance easy accessibility of information with adequate mechanisms to identify authorized users and ensure data integrity and confidentiality. A number of general precautions are herewith provided to minimize the possibility of virus infection.

1. Back up important data frequently and keep them in a safe place other than the computer. Set up a backup schedule and comply with it punctually.
2. Patch the operating system as quickly as possible to block the potential vulnerabilities that malware can exploit and sneak in.
3. Obtain the most recent virus definition to keep the antivirus application up-to-date.
4. Be suspicious of e-mail attachments from unknown sources and scan them first; be cautious when opening e-mail attachments even from known sources, because e-mail attachments are currently a major source of infection and sophisticated viruses can automatically send e-mail messages from other's address books.
5. Scan all new software before installing and opening, particularly the media that belong to other people. Sometimes even the trial and retail software has viruses.
6. Be extremely vigilant with external sources, such as CDs, diskettes, and Web links.
7. Always keep application's auto-protect running. Set the application default to auto-protection upon the launch of system.
8. Scan floppy disks at shutdown.

Additionally, many recent significant outbreaks of virus stem from Operating System (OS) vulnerabilities. From the perspective of the security community, many widespread security problems arguably might stem from bad interaction between humans and systems (Smith, 2003). Vulnerabilities can exist in large and complex software systems as well as human carelessness and sabotage. At least with today's software methods, techniques, and tools, it seems to be impossible to completely eliminate all flaws (Lindskog, 2000). Virus writers have demonstrated a growing tendency to exploit system vulnerabilities to propagate their malicious code (Trend-Micro, 2003). Operating systems consist of various and complex yet vulnerable software components that play a crucial role in the achievement of overall system security, since many protection mechanisms and facilities, such as authentication and access control, are provided by the operating system. Vulnerabilities and methods for closing them vary greatly from one operating system to another. Therefore, it is of vital importance to screen these following items in different OS to strive for greatest avoidance of malware attacks (Rash, 2004; SANS-Institute, 2003; Vijayan, 2004).

Microsoft Windows

- Internet Information Services (IIS)
- Microsoft SQL Server (MSSQL)
- Windows Authentication
- Internet Explorer (IE)
- Windows Remote Access Services
- Microsoft Data Access Components (MDAC)
- Windows Scripting Host (WSH)
- Microsoft Office Suite (Word and Excel)
- Microsoft Outlook and Outlook Express
- Windows Peer to Peer File Sharing (P2P)
- Simple Network Management Protocol (SNMP)
- Abstract Syntax Notation One (ASN.1) Library

Novell Netware

- NetWare Enterprise Server
- NetWare NFS
- Remote Web Administration Utility

Unix/Linux

- Open Secure Sockets Layer (SSL)
- Apache Web Server
- BIND Domain Name System (DNS) Server
- Remote Procedure Calls (RPC) Services
- Sendmail
- General UNIX Authentication Accounts with No Passwords or Weak Passwords
- Clear Text Services
- Simple Network Management Protocol (SNMP)
- Secure Shell (SSH)
- Misconfiguration of Enterprise Services NIS/ NFS

The traditional method of waiting for antivirus vendors to provide a strategy only *after* virus infections have occurred no longer fits with the needs of an enterprise, which should be more proactive to cope with the antivirus predicaments. In order to keep the environment virus-free, there are several combined recommended strategic stages that enterprises should follow (Carden, 1999; Rash, 2004; Rose, 1999; Smith, McKeen, & Staples, 2001):

1. **Method assessment:** Companies must first assess whether passive protection mechanisms (e.g., virus scanning and firewalls) are adequate for their needs or whether more active protection, such as vulnerability analysis and intrusion detection, is needed.
2. **Good User Education:** Always set security awareness high by informing users of the importance of a virus-free environment. An end-user training policy is necessary.
3. **Tight Control of User's Activation:** Redirect potentially harmful attachment or downloads, such as executables and macro-bearing documents, from untrustworthy sources. Different level users must comply with relevant policies.
4. **Information Encryption:** For high-profile or security-perceptive organizations, it's highly crucial to encrypt the information via the secure socket layer (SSL) embedded in the browsers or through Secure Electronic Transmission (SET). Encryption can eliminate essentially all hacker interception of the transmission itself.
5. **Internal Infrastructure Attention:** Managers need to ensure that the internal infrastructures, including firewalls to protect internal systems, are fully functioning and robust to attack. If they are not, greater attention must be placed in this area.
6. **Collect Vulnerability Information:** From internal and external sources, such as company IT teams, security vendor's Web page, third-party suggestions, and even hacker's information exchange webpage.
7. **Validate Accuracy of Information:** Check with a respected source and delete unrelated or unnecessary or out-of-date information.
8. **Form a Plan to Remediate Vulnerability:** Remediation includes applying the appropriate patches, changing hardware or application configurations, or making policy changes.
9. **Inventory the Environment:** Make sure you know what you have before patch it.
10. **Analyze Correlations Between the Assets and Vulnerability Knowledge:** Software tools may be able to help here.
11. **Fix the Problem then Check is Done Correctly.**

CONCLUSION

As mentioned above, the mushrooming growth of the Internet galvanizes and provides a hotbed for e-mail-borne macro viruses and worms. The mass-e-mailing worms, such as Mydoom, So-Big, and the recent Sasser and Netsky, outbreak like wildfire in a short period of time via the globally connected networks and pose a significant challenge to our cyber society prior to our awareness and recognition. According to Trend Micro's research, during the 2001 to 2003 time period, 100 percent of outbreaks had Internet worm-like characteristics and most worms use email and some form of social engineering to entice users to click and execute attachments (Trend-Micro, 2003). Therefore, an enterprise must protect itself and its employees from all of the risks associated with email by enforcing the right sort of policies and procedures. The policy for the email access has to be defined what "appropriate" content is for the business, characterizing all the rest as "inappropriate." A global policy for all employees of the company might not fit the entire working situation that different employees face. Different departments in an organization might need different policies, and different rules can be set for different user groups. Also, the policy must apply e-mail content analyzing and filtering controls by deploying attachment filter, anti-spam filter, message size filter, and content filter (Paliouras, 2002; Trend-Micro, 2002).

In the future, the sophistication of malware is expected to continue to grow, requiring ever-increasing sophistication of methods of malware prevention, detection, and remediation. It is estimated that in the near future, with newer viruses and worms which expose open ports (and don't require any recipient activity) and because millions of PCs are left on day and night, the time required to infect computers globally will be cut from days to minutes. Managers of large and small organizations must grow more vigilant in their efforts to prevent the costly damages resulting from malware vulnerability. Employees must be trained effectively, policies and procedures must be followed carefully, and new capabilities in the war on malware must be cultivated and implemented. The war will not end, but the outcomes can be influenced, and the damages can be reduced by proper managerial and technical awareness and action.

REFERENCES

Arce, I. (2004). More bang for the bug: An account of 2003's attack trends. *IEEE Security and Privacy, 2*(1), 66-68.

Caloyannides, M.A. (2003). Digital evidence and reasonable doubt. *IEEE Security and Privacy, 1*(6), 89-91.

Camp, S.V. (2004). Antivirus category remains healthy. *Brandweek, 45*(7), 14.

Carden, P. (1999). Antivirus software. *Network Computing, 10*(20), 78.

Gaudin, S. (2004). Virus attacks reach 'epidemic' proportions. *eSecurityPlant.com*

Goldsborough, R. (2003). A call to arms: How to stave off a computer virus. *Community College Week, 16*, 19. Cox Matthews and Associates Inc.

Harris, J. (2003). searchSecurity.com. Definitions online: *http://searchsecurity.techtarget.com/sDefinition/0%2C%2Csid14_gci213306%2C00.html*

Lindskog, S. (2000). *Observations on operating system security vulnerabilities.* Thesis for the degree of engineering (MS-PhD), Technical Report No 332L.

Paliouras, V. (2002). The need of e-mail content security. *Journal of Internet Security, 3*(1).

Palmer, C.C. (2004). Can we win the security game? *IEEE Security and Privacy, 2*(1), 10-12.

Rash, W. (2004). Disarming worms of mass destruction. *InfoWorld, 1*(2), 55-58.

Rash, W. (2004). What, me vulnerable? *InfoWorld, 1*(2), 57.

Roberts, P. (2003). Survey shows fewer, costiler viruses. InfoWorld Online: *http://www.infoworld.com/article/03/03/20/HNcostlier_1.html*

Rose, G., Huoy Khoo, & Straub, D.W. (1999). Current technical impediments to business-to-con-

sumer electronic commerce. *Communications of AIS, 1*(16).

Salkever, A. (2004). Mydoom's most damning dynamic. *Business Week Online* (pp. N.PAG): McGraw-Hill Companies, Inc. — Business Week Online.

SANS-Institute (2003). SANS Top 20 Vulnerabilities - The Experts Consensus. Online: *http://www.sans.org/top20/*

Smith, H.A., McKeen, J.D., & Staples, D.S. (2001). Risk management in information systems: Problems and potentials. *Communications of AIS, 7*(13).

Smith, S.W. (2003). Humans in the loop: Human-computer interaction and security. *IEEE Security & Privacy, 1*(3), 75-79.

SolutionsReview.com. (2003). How does antivirus work? Online: *http://www.solutionsreview.com/Antivirus_how_do_antivirus_software_work.asp*

Stein, A. (2004). Microsoft offers MyDoom reward. CNN/Money online: *http://money.cnn.com/2004/01/28/technology/mydoom_costs/*

Symantec (2004). What is the difference between viruses, worms, and Trojans? Online: *http://service1.symantec.com/SUPPORT/nav.nsf/pfdocs/1999041209131106*

Tan, J. (2004). 2003 viruses caused $55B damage. *ComputerWorld* online: *http://www.computerworld.com/securitytopics/security/story/0,10801,89138,00.html*

Trend-Micro (2002). E-mail content security management. *Trend Micro, Inc.*

Trend-Micro (2003). The trend of malware today: Annual virus round-up and 2004 forecast. *Trend Micro, Inc.*

Vijayan, J. (2004). Microsoft issues patches for three new Windows vulnerabilities. *ComputerWorld*.

KEY TERMS

Antivirus Software: A class of programs that searches networks, hard drives, floppy disks, and other data access devices, such as CD-ROM/DVD-ROM and zip drives, for any known or potential viruses. The market for this kind of program has expanded because of Internet growth and the increasing use of the Internet by businesses concerned about protecting their computer assets.

Live Update: An integrated/embedded program of the antivirus software. It is intended to provide frequent (weekly or self-scheduled) virus definition and program modules updates. It can automatically run in the background and connect ("talk to") the server of antivirus software vendor to identify whether an update is available. If so, it will automatically download the update and install the update.

Morphing Virus/PolymorphicVirus: They are undectable by virus detectors because they change their own code each time they infect a new computer and some of them change their code every few hours. A polymorphic virus is one that produces varied but operational copies of itself. A simple-minded, scan string-based virus scanner would not be able to reliably identify all variants of this sort of virus. One of the most sophisticated forms of polymorphism used so far is the "Mutation Engine" (MtE) which comes in the form of an object module. With the Mutation Engine any virus can be made polymorphic by adding certain calls to its assembler source code and linking to the mutation-engine and random-number generator modules. The advent of polymorphic viruses has rendered virus-scanning an ever more difficult and expensive endeavor; adding more and more scan strings to simple scanners will not adequately deal with these viruses.

Scanning (can be scheduled or batch)**:** The activity/performance launched by the antivirus software to examine the files and to inspect any malicious codes resided inside the files according to the software's definition files.

Stealth Virus: A virus that hides the modifications it has made in the file or boot record, usually by monitoring the system functions used by programs to read files or physical blocks from storage media, and forging the results of such system functions so that programs which try to read these areas see the original uninfected form of the file instead of the actual infected form. Thus the virus modifications go

undetected by antivirus programs. However, in order to do this, the virus must be resident in memory when the antivirus program is executed.

Virus Definition File (subscription service): A file that provides information to antivirus software to find and repair viruses. The definition files tell the scanner what to look for to spot viruses in infected files. Most scanners use separate files in this manner instead of encoding the virus patterns into the software, to enable easy updating.

Virus Signature: A unique string of bits, or the binary pattern, of a virus. The virus signature is like a fingerprint in that it can be used to detect and identify specific viruses. Antivirus software uses the virus signature to scan for the presence of malicious code.

Measuring the Potential for IT Convergence at Macro Level

Margherita Pagani
Bocconi University, Italy

WHAT IS CONVERGENCE?

Convergence describes a process change in industry structures that combines markets through technological and economic dimensions to meet merging consumer needs.

It occurs either through competitive substitution or through the complementary merging of products or services, or both at once (Greenstein & Khanna, 1997).

The main issues in the process of convergence have been investigated in the literature (Bradley, Hausman and Nolan, 1993; Collins, Bane and Bradley, 1997; Yoffie, 1997; Valdani, 1997, 2000; Ancarani, 1999; Pagani, 2000).

The numerous innovations that could lead to convergence between TV and online services occur in various dimensions.

The *technology dimension* refers to the diffusion of technological innovations into various industries. The growing integration of functions into formerly separate products or services, or the emergence of hybrid products with new functions is enabled primarily through digitization and data compression. Customers and media companies are confronted with technology-driven innovations in the area of transport media as well as new devices. Typical characteristics of these technologies are digital storage and transmission of content, and a higher degree of interactivity (Schreiber, 1997; Rawolle & Hess, 2000).

The *needs dimension* refers to the functional basis of convergence: Functions fulfill needs of customers, which can also merge and develop from different areas. This depends on the customers' willingness to accept new forms of need fulfillment or new products to fulfill old needs.

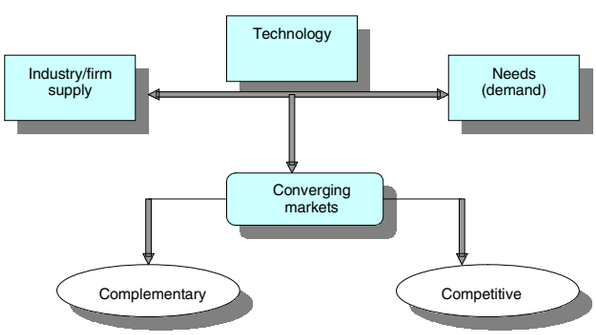
This dimension in the process of convergence refers to the formation of integrated and convergent 'cluster of needs' (Ancarani, 1999) that is the tendency of customers to favour a single supplier for a set of related needs (Vicari, 1989).

The *competitive dimension* refers to mergers, acquisitions, alliances and other forms of cooperation – often made possible by deregulation – among operators at different levels of the multimedia value chain. Competitive dynamics influence the structures of industries just as it does the typical managerial creativity of the single factory in originating products and services, combining know-how to create new solutions and removing the barriers among different users' segments.

A strategic intent is at play on the part of enterprises to use the leverage of their own resources within a framework of incremental strategic management in order to deploy them over an ever-increasing number of sectors.

One thinks, in this regard, of Hamel's (1996, 2000) concept of 'driving convergence'. This concept places the firm and its own competitive strategies at the control of the process of industry convergence.

Figure 1. A summary of dimensions and basic forms of convergence



Source: Adapted from Dowling, Lechner, & Thielmann, 2000

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

In general, the concept of digital convergence is used to refer to three possible axes of alignment (Flynn, 2000).

- convergence of devices
- convergence of networks
- convergence of content

Although there is evidence in digital environments of limited alignment in some of these areas, there are considerable physical, technical, and consumer barriers in all three areas.

CONSTRAINTS TO CONVERGENCE

There are three different types of constraints on the convergence of the devices that are used to access the three digital platforms (digital TV, personal computers and mobile devices). These constraints can be summed up in the form of the three following questions.

1. Is it physically possible to merge the two devices?
2. Is it technically possible to merge the two devices?
3. Will consumers want to use the merged device?

Given that we are talking about three different types of network-access devices here (TV, PCs, and mobile devices), there are three potential areas of convergence: PC and TV (web TV or Internet access from digital TV), PC and mobile phones (mobile television), and TV and mobile phone.

In the physical domain, the barriers to PC and TV convergence lie principally with respect to the size of the input device and its portability. The barriers to PC and mobile phone convergence in the physical domain are rather more acute, and there is a divergence along every physical measure (size of display device, size of input device, and portability).

Technical requirements either affect the available transport media, the addressed end device, or both. Three important aspects dominate in this area.

- the access mechanism
- the number of simultaneous recipients
- the support of feedback channels in the case of transmission media

With regard to the access mechanism, a distinction between push and pull mechanisms must be made. Pull-oriented access is characterized by the data transmission being triggered by the end user (which is typical for Web applications or video on demand), whereas push-oriented transmission is triggered by the sender. Push services can be time scheduled (e.g., television broadcast).

Device-specific requirements mainly affect reproduction, storage capabilities, and input facilities. Displaying and synchronizing different kinds of media types is a basic demand with regard to reproduction. A distinction between static (time invariant as text, graphics, and pictures) and dynamic (time variant as video and audio) media types has to be made (Grauer & Merten, 1996). Next, storage capabilities enable synchronous download and consumption of contents in the case of online media usage. Typically, end devices with roots in information technology (like PCs, PDAs [personal digital assistants], and notebooks) possess sufficient, persistent storage capacity. In contrast, most of the entertainment electronics lack comparable characteristics.

Another important aspect of end devices is input facilities. Typically, PC-based end devices possess the most advanced mechanisms for user input (keyboard, mouse, joystick, etc.). In contrast, mobile or TV-based devices usually lack sophisticated input facilities.

Table 1. A summary of physical characteristics of consumer devices

Characteristic	TV	PC	Mobile phone
Size of display device	Large	Large	Small
Size of input device	Small	Large (keyboard)	Small (keypad)
Portability	Low	Medium	High

Table 2. A summary of technical characteristics of consumer devices

Characteristic	TV	PC	Mobile phone
Display type	Cathode ray tube	Cathode ray tube	Liquid crystal display
Display resolution	Medium	High	Low
Display scanning mode	Interlaced	Progressive	Progressive
Display refresh rate	Medium	High	High
Processing power	Low	High	Low
Storage	Low	High	Low
Power requirement	High	High	Low

A comparison among the relevant technical characteristics of the three different types of consumer devices (Table 2) shows that there is little evidence of TV and mobile phone convergence as yet, and in any case, the technical constraints with respect to this particular combination are implicit in the consideration of the other instances.

Consumer attitudes (Noelle & Neumann, 1999) to devices that inhabit the TV environment as opposed to the PC and mobile telephony environments are also widely different (Table 3).

End users have certain usage patterns and behaviors that are closely correlated to end devices and transport media. PC usage differs from TV usage in terms of user activity (active vs. passive) and purpose (information vs. entertainment). Another important aspect has to be considered in view of user attention.

The types of content that are carried over the PC and Internet, broadcast, and telephony networks show some sharply differentiated characteristics, and consumer usage and distribution differs across platforms (Table 4).

The ability to merge data about consumer preferences and transactional profiles across platforms is critical for any interactive media business, and this can be achieved through a process of cross-platform tracking.

A DEFINITION BASED ON PLATFORM PENETRATION AND CRM POTENTIAL

Customer relationship management (CRM) can be described as the process of attracting, retaining, and capitalising on customers. CRM defines the space where the company interacts with the customers. At the heart of CRM lies the objective to deliver a consistently differentiated and personalised customer experience, regardless of the interaction channel (Flynn, 2000).

Table 3. Differing consumer expectations for different platforms

Consumer expectations in TV space	Consumer expectations in PC space	Consumer expectations in mobile phone space
Medium, stable pricing of goods	High, unstable pricing of goods	Low, unstable pricing of goods
Infrequent purchase (once every 7 to 11 years)	Frequent purchase (every 18 months to 3 years)	Frequent purchase (every 18 months to 3 years)
Little requirement for software and peripheral upgrades	High requirement for software and peripheral upgrades	Medium requirement for software and peripheral upgrades
Works perfectly first time	Probably will not work perfectly first time	Probably will work first time
No boot-up time	Long boot-up time	No boot-up time
Low maintenance	High maintenance	Low maintenance
Low user intervention	High user intervention	High user intervention
Little or no technical support required	Substantial technical support required	Little technical support required

Table 4. A summary of content characteristics of the three major digital platforms

TV/broadcast content attributes	PC/Internet content attributes	Mobile telephony content attributes
Video heavy (moving pictures lie at its core, rather than text)	Video light (text and graphics lie at its core, rather than video)	Voice based (audio lies at its core, rather than text, graphics or video)
Information medium (the factual information transmitted is not very dense)	Information heavy (the factual information transmitted is dense)	Where non voice based material is transmitted, it is information light (any textual information transmitted in an SMS – Short Messaging Service or on a WAP - Wireless Application Protocol phone is sparse)
Entertainment based (to provide a leisure activity rather than learning environment)	Work based (to provide work related or educational information or to enhance productivity rather than to be entertained)	Both work based and socially based (to provide work related information or to enhance productivity rather than to be entertained).
Designed for social or family access	Designed to be accessed by solitary individuals	Designed to be accessed by two individuals
Centrally generated (by the service provider)	Both centrally generated (content on a CD-Rom or website) and user generated (email, chat, personalization, etc)	Predominantly user generated
User unable to influence content flow which is passively received rather than interacted with and linear in form	User typically interacts with the content producing a non linear experience	Where centrally generated content is provided, user typically interacts with the content.
Long form (the typical program unit is 25 minutes long)	Short form (video information tends to be in the form of clips or excerpts).	Short form (text and websites highly abbreviated, audio in form of clips or excerpts).

The business potential in x-media commerce is in attracting, retaining, and capitalising on customer relationships through interactive media channels. This suggests a definition of convergence based not on the merging of digital devices, networks, or content, but on the extent to which the transition to two-way digital networks facilitates consumer convergence or cross-platform customer relationship management.

From a CRM perspective, technology-based convergence taking place between different platforms is not a central concern. The key is that these different, often incompatible, technology platforms enable customers to interact with companies trough different channels, allowing those companies to increase the number of potential contact points with their customers.

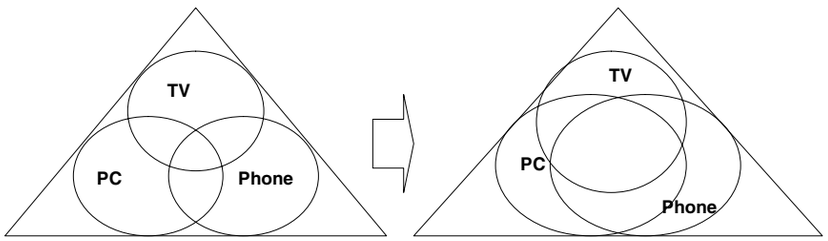
For media companies looking at their digital-investment strategy in a specific country and seeking to maximize their benefit from this type of convergence, it is key to know which territories exhibit the best potential for development so that those companies can decide where initially to test and/or introduce interactive applications or how to assess the likely success of existing projects in a CRM context.

The goal of the following model is to provide a methodology for convergence measurement.

The following three indicators for the measurement of convergence potential are considered.

- critical digital-mass index
- convergence factor
- interactivity factor

Figure 2.The effect of platform overlap



The Concept of Critical Digital-Mass Index

One cornerstone in the measurement of convergence potential is the extent to which digital platforms (such as digital TV, PCs and Internet access, and mobile devices) are present in a specific country. This will obviously make it easier to reap the efficiencies and economies of scale that CRM offers.

However, since CRM strategies derive their greatest benefits across multiple channels, one needs to measure the penetration of such platforms in combination. This combined measure (penetration of platform A plus penetration of platform B plus penetration of platform C) indicates the critical digital mass of consumers in any given territory.

The critical digital-mass index for a territory is created by adding together the digital TV penetration, mobile phone penetration, and PC and Internet penetration in each territory.

(Penetration of digital TV) + (Penetration of mobile telephony) + (Penetration of PC and Internet)

The Convergence Factor

The potential for CRM is greatest where the same consumers are present across all three digital platforms: This would be the optimal situation for an integrated multichannel CRM strategy. The degree of overlap tends to be much higher when overall digital penetration is higher (this is not a linear relationship). If penetration of digital TV, PC and Internet access, and mobile telephony are all above 50%, the number of consumers present across all three is likely to be much more than 5 times greater than is the case if penetration is at only around 10% in each case. Figure 2 illustrates this effect (The area within each triangle represents the boundaries of the total consumer universe.).

This means that the critical digital-mass indicator needs to be adjusted upward for higher overall penetration levels.

Applying simple probability theory, we give a way of measuring the rate at which cross-platform populations increase as penetration of those platforms increases. We assume that the three events of having all three devices are independent.

The convergence factor is derived from the penetrations of the three platforms multiplied by each other.

(Penetration of digital TV/100) x (Penetration of mobile telephony/100) x (Penetration of PC and Internet/100) x 100

In order to give some explanation of how the formula has been derived, we can suppose that the penetration of platform A is 10%. According to simple probability theory, the likelihood that one person chosen at random from the population is a member of that platform is 10:1. If penetration of platform B is also 10%, then the likelihood that the person we have chosen at random will also be a member of platform B is 100:1. If penetration of platform C is, again, 10%, then the odds that our initially chosen person is on that platform, too, is 1,000:1. In a population of 1 million individuals, in other words, the chances are that there are 1,000 people who fall into this category.

Take the opposite end of the penetration case, however, and assume that platforms A, B, and C all have a penetration level of 90%. Using the same methodology, in a population of 1 million, the chances are that just over 7 out of 10 people are on all three platforms (0.9x0.9x0.9); in a population of 1 million, this is equivalent to 729,000 people. Finally, of course, when all three platforms reach universal penetration, everyone in the population is a member of all three (that is, the probability is 1).

It is likely that the relationship between members of different platforms is not completely random in this way. For instance, early adopters tend to buy in early to all new technologies, and there is known to be higher PC penetration in digital-TV homes (presumably an income-related effect). So this way of assessing CRM potential probably somewhat underestimates the reality, at least in the early stages of an evolving digital market. However, by and large, the digital TV, PC and Internet, and mobile telephony markets surveyed have moved out of the early-adopter phase.

The Relevance of Interactivity

Another important element in assessing CRM potential is the extent to which the digital networks facilitate customer tracking.

Four levels of interactivity are considered.

- local
- one way
- two-way low
- two-way high

Networks exhibiting a high level of two-way interactivity are obviously those where CRM potential is greatest. In general, digital TV networks offer a lower level of interactivity than mobile and PC-Internet ones. For this reason in the following formula we assume different weights. Considering digital TV, we need to also distinguish the different interactivity level related to the specific transmission system (satellite, terrestrial, optical fiber, ADSL).

The interactivity factor for a territory is calculated according to the following formula.

((Penetration of digital TV) + (Penetration of mobile telephony x 2) + (Penetration of PC and Internet x 2))/5

The Convergence Index

The convergence index is generated as follows.

*[Critical digital-mass index * (1 + Convergence factor) * (1 + Interactivity factor)]*
*[(D + M + I) * (1 + D * M * I) * (D + 2M + 2I)/ 5)]*

D = Digital TV penetration, M = Mobile telephony penetration, I = PC-Internet penetration

This index represents the critical digital mass of consumers. It is possible to derive estimates of the number of consumers likely to be present across all three platforms by the simple expedient of taking the population of each territory and multiplying it by the triple-platform penetration factor. It is also possible to give an indication of the number of consumers likely to be present across two platforms by doing a double-platform penetration calculation.

CONCLUSION

The conclusions that the model generates are designed to give companies guidance as to how the broad convergence picture will evolve over time in each country studied. The goal of this model is not to obtain the accurate size of these cross-platform populations. This model is also a good starting point to address other related questions and allows for a number of further research to profile some of the key players in each territory and channel in order to assess which types of companies are best placed to exploit this newly defined type of convergence. For companies looking at their digital-investment strategy and seeking to maximize their benefits from consumer convergence, it is key to know which territories exhibit the best potential for development. Companies knowing this can decide where initially to test or introduce CRM systems, or how to assess

the likely success of existing projects in a CRM context. Further research is need is to integrate this model with marketing issues in order to consider the *intensity* of the use and also the *kind of use.* The CRM potential is much more attractive if users employ more than an IT channels for the same final purpose (e.g. job either entertainment). If uses are very heterogeneous, concerns such as intensity, individual VS collective use and coherence, the cross CRM potential should be much less appealing than in the opposite case.

The model assesses if there are enabling conditions for cross CRM and if these conditions are better in one country or in another.

REFERENCES

Ancarani, F. (1999). *Concorrenza e analisi competitiva.* EGEA. Milan, Italy.

Bradley, S., Hausman, J., & Nolan, R. (1993). *Globalization, technology and competition: The fusion of computers and telecommunications in the 1990s.* Boston: Harvard Business School Press.

Brown, S. L., & Eisenhardt, F. M. (1999). *Competing on the edge.* Boston: Harvard Business School Press.

Collins, D. J., Bane, W., & Bradley, S. P. (1997). Industry structure in the converging world of telecommunications computing and entertainment. In D. B. Yoffie (Ed.), *Competing in the age of digital convergence,* 159-201. Boston, MA: Harvard Business School Press.

Dowling, M., Lechner, C., & Thielmann, B. (1998). Convergence: Innovation and change of market structures between television and online services. *Electronic Markets Journal, 8*(4, S), 31-35.

Flynn, B. (2001). Digital TV, Internet & mobile convergence. *Report Digiscope.* London, UK: Phillips Global Media.

Gilder, G. (2000). *Telecoms.* New York: Free Press.

Grant, A. E., & Shamp, S. A. (1997). Will TV and PCs converge? Point and counter point. *New Telecom Quarterly*, 31-37.

Greenstein, S., & Khanna, T. (1997). What does convergence mean? In D. B. Yoffie (Ed.), *Competing in the age of digital convergence.* Boston, MA: Harvard Business School Press.

Grauer, M. & Hess, T. (2000). New digital media and devices: An analysis for the media industry. *Journal of Media Management, 2*(2), 89-98.

Noelle-Neumann, E., Shultz, W. & Wilke, J. (1999). *Publizistic Massenkommunikation,* Frankfurt, Germany, A.M., Fischer.

Owen, B. M. (1999). *The international challenge to television.* Boston: Harvard University Press.

Pagani, M. (2003). Measuring the potential for IT convergence at macro level: A definition based on platform penetration and CRM potential. In C. K. Davis (Ed.), *Technologies and methodologies for evaluating information technology in business,* 123-132. Hershey, PA: Idea Group Publishing.

Pine, B. J., & Gilmore, J. M. (1999). *The experience economy.* Boston: Harvard Business School Press.

Rawolle, J., & Hess, T. (2000). New digital media and devices: An analysis for the media industry. *Journal of Media Management, 2*(2), 89-98.

Schreiber, G.A. (1997). *Neue Wege des Publizierens.* Wiesbaden: Vieweg.

Valdani, E., Ancarani, F., & Castaldo, S. (2001). *Convergenza: Nuove traiettorie per la competizione, 3,* 89-93. Milan: Egea.

Vicari, S. (1989). *Nuove dimensioni della concorrenza.* Milano EGEA.

Yoffie, D. B. (1997). *Competing in the age of digital convergence.* Boston: Harvard Business School Press.

KEY TERMS

Convergence: The term describes a process change in industry structures that combines markets through technological and economic dimensions to meet merging consumer needs. It occurs either through competitive substitution or through the complementary merging of products or services, or

both at once. In general, the concept of digital convergence is used to refer to three possible axes of alignment.

- convergence of devices
- convergence of networks
- convergence of content

Convergence Factor: It measures the rate at which cross-platform populations increase as penetration of platforms increases. The convergence factor is derived from the penetrations of the three platforms multiplied by each other.

Convergence Index: This index represents the critical digital mass of consumers, and it estimates the number of consumers likely to be present across all three platforms by the simple expedient of taking the population of each territory and multiplying it by the triple-platform penetration factor.

Critical Digital-Mass Index: It measures the extent to which digital platforms (digital TV, PC, Internet access, and mobile phones) are present in a given territory. It is created for a territory by adding together the digital TV penetration, mobile phone penetration, and PC Internet penetration.

CRM (Customer Relationship Management): This can be described as the process of attracting, retaining, and capitalising on customers. CRM defines the space where the company interacts with the customers. At the heart of CRM lies the objective to deliver a consistently differentiated and personalised customer experience, regardless of the interaction channel.

X-Media: The opporunity to transmit the digital content through more than one type of media (television, Internet, wireless).

Message-Based Service in Taiwan

Maria Ruey-Yuan Lee
Shih-Chien University, Taiwan

Feng Yu Pai
Shih-Chien University, Taiwan

INTRODUCTION

The number of cellular phone subscribers has increased 107% in Taiwan, based on the Directorate General of Telecommunications reports (http://www.dgt.gov.tw/flash/index.html). Meanwhile, Internet users have reached a total of 8.8 million, and mobile Internet users have broken the record of 3 million in 2004. The combination of information and telecommunication technologies has brought people a new communication method—cellular value-added services, which have become a lucrative business for telecommunication providers in Taiwan.

One result of the cellular value-added services presented to the public, which brought the information-based, messaging-based and financial services into one kit, was that people not only could communicate through the cellular phones, but they also could use them as versatile handsets. DoCoMo, a famous Japanese telecommunication provider, has successfully cultivated the cellular value-added services. Its success lies in two areas: (1) content and Web site providers are willing to share their technical support; and (2) an automated payment system was established to assist cash flow between providers and even beef up the whole industry by associating related business partners (Natsuno, 2001).

In addition to DoCoMo's case, the telecommunication service providers in Taiwan have provided various cellular value-added services. However, the popularity of the service did not turn out to be as good as expected. We are wondering why. Telecommunication providers began to adjust the fee of short message service (SMS) down to 25% maximally since June 2004 in Taiwan. The idea of lowering fees is to stimulate the popularity of SMS usage. Would that bring a collateral effect to the providers of cellular value-added service positively or negatively? Therefore, this research will discuss the challenges facing Taiwanese cellular value-added service providers. Hinet, Taiwan Cellular Corporation (TCC), and Flyma (online service providers) have been chosen as research companies.

BACKGROUND

The great innovation of Information Technology (IT) has brought both cellular phone and Internet technology to a reality; a high penetration rate of cellular phone subscribers and a great popularity of the Internet has completely changed communications among people. With these two new technologies, people can communicate with each other without a concern about when and where. The created value of cellular value-added service has been considered as a significant issue in this research.

According to the Marketing Intelligence Center, the cellular value-added service can be categorized as message-based service, entertainment service, financial service, and information service. Table 1 shows the categorization.

This research focuses mainly on message-based services (short message service(SMS), e-mail, and Multi-Media Service (MMS). Three different types of cellular value-added service industries have been chosen as case studies, including a System Service Provider (Taiwan Cellular Corporation), an Internet Service Provider (ISP) (Hinet), and a Value-Added Service Provider (Flyma). Taiwan Cellular Corporation (TCC) (www.tcc.net.tw) is one of the biggest telecommunication providers in Taiwan. It specializes in network infrastructure, product offering, technology development, and customer services. The value-added services in TCC include SMS, MMS, entertainment, and so forth. HiNet (www.hinet.net) is Taiwan's largest ISP and has, by far, the largest number of users in Taiwan. Hinet's value-added

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. The cellular value-added service categorization (marketing intelligence center, http://mic.iii.org.tw/index.asp)

Category	Description	Application
Message-based service	Providing users real-time message services	Short message service(SMS), e-mail, multi-media service
Entertainment service	Providing users recreational services	Downloading hot pictures, music tunes, and games
Financial service	Providing users services of financial issues	Mobile banking, mobile shopping, etc.
Information service	Providing users up-to-date information	Information of weather, news, sports, mapping, etc.

service includes voice over IP (VOIP), games, MMS, and so forth. Flyma (www.flyma.net) is a small enterprise company specializing in wireless value-added services such as MMS, e-mail, and so forth.

The perspective analysis is based on the Balance Scorecard (BSC) (Kaplan & Norton, 1992, 1993, 1996a,b,c,d). BSC has been used as a strategic management system and performance measurement. The BSC suggests that we view the organization from four perspectives and to develop metrics, collect data, and analyze them relative to each of these perspectives: the learning and growth perspective, the business process perspective, the customer perspective, and the financial perspective. In this article, we customize the BSC's four perspectives to be cellular value-added service: Service Charging, Customer Relationship, Business Partnership, Innovation and Learning. We illustrate the four perspectives' relations with each company's vision and business strategies.

PERSPECTIVE ANALYSIS

Based on the four fundamental perspectives of BSC, we customize the perspectives to be the following cellular value-added service perspectives.

1. **Service Charging:** Due to the company's internal financial confidential information, we focus mainly on the charging comprisal of SMS fee.
2. **Customer Relationship:** We focus on the customer segmentation of each case and the process of CRM.
3. **Business Partnership:** We discuss the relationship between these related business partners.
4. **Innovation and Learning:** We compare the program of human resource enhancement for each case.

Figure 1 shows the proposed cellular value-added service perspectives with their relationships.

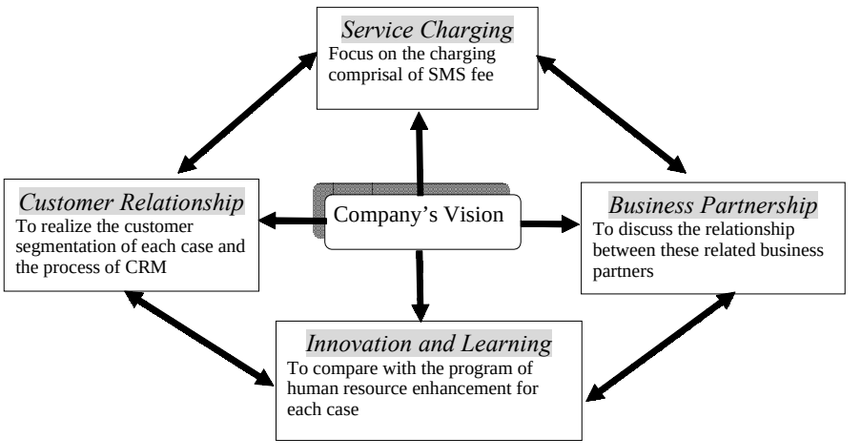
Service Charging

The service charge consists of an MSN fee per cost, an estimated production fee, and an access fee. The access fee is the administration fee to the ISP provider. The access fee charge is 20% of MSN per cost minus the production fee. In other words, only the value-added service provider (Flyma) needs to pay ISP providers, whereas both Hinet and TCC do not need to pay the access fee, because they are the ISP. Table 2 shows MSN service charging structure in Taiwan. The figure is shown in NT dollars.

Customer Relationship

Due to the saturated market of cellular phone subscribers, telecom providers begin to provide subscribers cellular value-added service. Based on the completed network of information infrastructure, Hinet successfully brought its services into each subscriber's home, which resulted in a good customer relationship. Moreover, TCC has considered the high quality of customer service as the company vision and has established a customer service department as an

Figure 1. A proposed cellular value-added service perspectives with their relationships


Service Charging
Focus on the charging comprisal of SMS fee
Customer Relationship
To realize the customer segmentation of each case and the process of CRM
Company's Vision
Business Partnership
To discuss the relationship between these related business partners
Innovation and Learning
To compare with the program of human resource enhancement for each case

individual company. The scale of customer service is much smaller then the two former companies; Flyma cannot afford a large budget for customer service. However, it still offers and maintains excellent customer support through the Web site.

However, apart from the services, these companies need to consider what else can affect the customer relationship. For example, the falsities through message-based service of cellular phones have rampantly come out and seriously damage the property of customers. Thus, customers usually have a negative impression of the message-based service in Taiwan.

Business Partnership

This perspective refers to business partnerships between companies. The partnership can be categorized as digital content providers, technical support companies, and affiliated business. Table 3 summarizes these companies' business partnerships. Generally speaking, Hinet offers a stable digital content provider and system maintenance. By introducing the Japanese production, TCC can increase the technology of manufacturing digital content. Flyma concentrates on the Web site design and provides an online customer service.

Innovation and Learning

This perspective includes employee training and corporate cultural attitudes related to both individual and corporate self-improvement. Based on the interview results, we found that Hinet possessed the most institutionalized system of human resource enhancement. Flyma pays much attention to the enhancement program; for example, it appoints a team to Japan for experiencing manufacturing digital content, whereas TCC particularly highlights the enhancement of customer service quality and provides its employees with knowledge of the relationship maintenance, customer relationship, and so forth.

ANALYZING RESULTS

The BSC's four perspectives provide a clear description as to what companies should measure in order to balance the perspective. We also recognize some

Table 2. A comparative figure of MSN service charge in Taiwan

Company	MSN per cost	Production Fee	Access fee	Net
Hinet	2	1.0	none	1.0
TCC	2.5	1	none	1.5
Flyma	2	1	1 ((2-1)*20%)	0.8

Table 3. Business partnership for value-added services

	Digital Content Provider	Technical Support Company	Affiliated Business
Hinet	Cooperating with other digital content providers and its research lab	Cooperating with a Canadian network solution company, Nortel Networks Corp.	Affiliating with the domestic Web site
Taiwan Cellular Corporation	Cooperating with Japanese company and introducing an up-to-date picture and tune	Granted to its subsidiary company, Howin Corp.	Affiliating with the domestic Web site
Flyma	Cooperating with the Japanese company, Ricoh	Cooperating with domestic telecom providers	Affiliating with education institution and government

of the weaknesses and strengths of the three value-added service companies in Taiwan.

Service Charging

In order to stimulate the frequency of cellular phone message usage, telecommunication providers have decreased the service charge of short message services since June 2004. In this situation, cost-down of cellular value-added service and good business partnership are very important for providers to maintain their advantage. A cheaper service charge and a better, more practical cellular value-added service can encourage customers to purchase the service.

Customer Relationship

- **Emphasize the Quality of Customer Service:** Speaking of the customized market, customers are no longer looking for quantity but also quality of service. Providing a nice quality of customer service is the fundamental business strategy in the recent market.
- **Emphasize the Quality of Cellular Value-Added Service:** Cellular value-added service providers only think about how to increase the number of cellular phone subscribers in order to increase market share. Price-cutting is the most typical promotional method in the recent saturated cellular phone market. However, the outcome of this promotion is not as good as expected. Consider an analogy of farming and this promotion. When a farmer wishes to increase his benefit from farming, a high quality of cultivation is more important than the quantity. Therefore, the cellular value-added service providers should pay more attention to how to improve their quality of cellular value-added service.

Business Partnership

Because of the great support from the government, Taiwanese digital content providers can concentrate on providing more practical digital content in order to enrich the value of cellular value-added service and to stimulate the popularity. In addition, technical support providers need to provide stable systems and also focus on integrating the various working systems from each provider, such as the payment system. Furthermore, in order to increase the usage of cellular value-added service, the application layer is what Web site providers need to induce more cross-industry companies to join and to make use of cellular value-added service, eventually as a national routine.

Innovation and Learning

Compared with the advanced digital content industry in Japan and Korea, the industry in Taiwan is still lagging far behind. Through importing digital content from Japan and Korea, Taiwan can provide customers with more options and also bring the Taiwanese digital content providers more ideas of the production base on technical support. Furthermore, the Taiwanese cellular value-added service providers also can appoint a team to Japan and Korea in order to experience a developed working environment and then come up with a positive effect technically and mentally for each team member.

DISCUSSION

The reason for the unpopularity of cellular value-added services in Taiwan could be clarified as follows:

1. **The Demand Rarely Reaches Economies of Scale:** In 2002, the short message service usage volume was around 2.1 billion, according to the DGT; the following year, there was a 14.3% growth rate. Even if there is a nice growth rate, demand rarely reaches economics of scale, according to Chunghwa Telecom, a famous Taiwanese telecom provider. Thus, it rarely grants a cost-down on cellular value-added service.
2. **The High Service Charge:** Compared with the 40 cents per short message service fee in mainland China, according to the consumer's foundation, the service charge still remains at a high price in Taiwan. Although the telecom providers have adjusted down to 25% maximally, the result hardly encourages the willingness of customer usage.
3. **Low Functional Support of Cellular Phone Models:** In order to carry out the practical and diversified content of cellular value-added service into reality, it is necessary to match with new models of cellular phones. However, it is difficult to promote a new handset at the same price as an old model in the short term. Pricing is still considered the most important issue for customers. Due to the low rates of cellular phone repurchasing, customers still use old cellular phones and only partly use the new cellular value-added service.
4. **Distorting the Positive Usage of Mobile Value-Added Service:** Recently, falsity through message-based service of cellular phones rampantly came out and seriously harmed the property of customers. Thus, customers mostly have a negative impression of the message-based service.

CONCLUSION

Based on the BSC theory, we aimed to analyze three different types of cellular value-added service companies in Taiwan. The research involved conducting field interviews. We have viewed the companies from four perspectives, developed metrics, collected data, and analyzed them relative to each of these perspectives: Service Charging, Customer Relationship, Business Partnership, and Innovation and Learning.

Based on the analysis, we have found that for the sake of small user numbers compared with telecom companies, online service providers need to provide diversified contents of value-added services in order to increase the company profile. Both telecom companies and online service providers need to concentrate on maintaining a high quality of consumer service in a long-term business strategy. Considering the saturation of cellular phone users, both telecom companies and online service providers should enhance the quality of value-added services instead of the quantity of cellular phone users. Inviting cross-industry business partners is to create a mobile value-added service environment and make use of mobile value-added services as a nationwide action.

To maintain future trends of cellular value-added services in Taiwan, we suggest the following:

1. **Beef up the Whole Industry.** The most important issue for cellular value-added service providers is to create a win-win business with their business partners. We suggest implementing a high quality of value-added service at a competitive price and having a systemic cooperation with related business partners such as digital content providers and cellular phone makers.
2. **Stimulate the Market.** How to build up an e-society still remains a big concern for cellular

value-added service providers. The suggestion is that providers need to bring up different industries, such as financial, computing, and entertainment, into one platform, which will enhance customers' convenience of information searching. Consequently, a higher penetration of cellular value-added service will result with greater popularity.

3. **Dissolve the Negative Impression.** Recently, the falsity through message-based service of cellular phones rampantly came out and made people have a negative impression of it. Thus, the providers need to lift the qualification bar of inspecting mass message senders.

For future work, we would like to enlarge the research scale. This research only focused on the Taiwanese cellular value-added service providers. In order to provide a more impartial and wider thinking of business running to Taiwanese providers, it is worth researching the business model in countries with successful cases, such as Japan and Korea.

ACKNOWLEDGEMENT

We would like to thank Yaw-Tsong Chen, the CEO of Flyma City Corp., for his great support and help during this research.

REFERENCES

Kaplan, R.S., & Norton, D. (1992). The balanced scorecard—Measures that drive performance. *Harvard Business Review, 70*(1), 71-79.

Kaplan, R.S., & Norton, D. (1993). Putting the balanced scorecard to work. *Harvard Business Review, 71*(5), 134-147.

Kaplan, R. S., & Norton, D. (1996a). Using the balanced scorecard as a strategic management system. *Harvard Business Review, 74*(1), 75-85.

Kaplan, R.S., & Norton, D. (1996b). The balanced scorecard: Translating strategy into action. Boston, MA: Harvard Business School Press.

Kaplan, R.S., & Norton, D. (1996c). Link the balanced scorecard to strategy. *California Management Review, 39*(1), 53-79.

Kaplan, R.S., & Norton, D. (1996d). Strategic learning and the balanced scorecard. *Strategy & Leadership, 24*(5), 18-24.

Natsuno, T. (2001). *I-mode strategy*. West Sussex, Chichester, UK: John Wiley & Sons.

KEY TERMS

Balanced Scorecard (BSC): A strategic management system and performance measurement.

Cellular Value-Added Service Categories: Cellular value-added service can be categorized as message-based service, entertainment service, financial service, and information service.

Entertainment Service: Providing users recreational services (e.g., downloading hot pictures, music tunes, and games).

Financial Service: Providing users services of financial issue (e.g., mobile banking, mobile shopping, etc.).

Information Service: Providing users up-to-date information (e.g., information of weather, news, sports, mapping, etc.).

Message-Based Service: Providing users real-time message services (e.g., short message service (SMS), e-mail, and multi-media service).

Methods of Research in Virtual Communities

Stefano Pace
Bocconi University, Italy

INTRODUCTION

The Internet has developed from an informative medium to a social environment where people meet together, exchanging messages and emotions and establishing friendships and social relationships. While the Internet was originally conceived as a commercial marketspace (Rayport & Sviokla, 1994) with new opportunities for both firms and customers (Alba, Lynch, Weitz, Janiszevski, Lutz, Sawyer, & Wood, 1997), nowadays the social side of the Web is a central phenomenon to truly understand the Internet. Social gratification is among the most relevant motivations to go online (Stafford & Stafford, 2001). People socialise through the Internet, adding a third motivation to their online activity, other that the pleasure of surfing in itself (the "flow experience" described by Hoffman and Novak, 1996) and the usefulness of finding information.

Virtual communities are springing both as spontaneous aggregation (like the Usenet newsgroups) or forums promoted and organised by Web sites. The topics of a community range from support for a disease to passion for a given product or brand. The intensity and relevance of the virtual sociality cannot be discarded. Companies can receive useful and actionable knowledge around their own offer studying the communities devoted to their brand. Hence social research should adopt refined tools to study the communities achieving reliable results. The aim of this work is to illustrate the main research methods viable for virtual communities, examining pros and cons of them.

VIRTUAL COMMUNITIES

A virtual community can be defined as a social aggregation that springs out when enough people engage in public conversations, establishing solid social tie (Rheingold, 2000). The study of virtual communities has increased following the development of the phenomenon. One of the first works is that of Rheingold that studied a seminal computer conferencing system: "The Well" ("Whole Earth 'Lectronic Link"). Starting from that, lots of researchers have deepened different facets of the Internet sociability. The methods employed are various: network analysis (Smith, 1999), actual participation as ethnographer in a virtual community (Kozinets, 2002), documentary and content analysis (Donath, 1999), interviews (Roversi, 2001), and surveys (Barry, 2001).

The aims of these studies can be divided into two main and intertwined areas: sociological and business-based. The former is well understandable, due to the relevance that the virtual sociality has gained today. Turkle (1995) uses the expression "life on the screen" to signal the richness of interactions available in the web environment; Castells (1996) reverses the usual expression "virtual reality" into "real virtuality", since the virtual environment cannot be considered a sort deprivation from life, but one of its enhancements and extensions. Regarding the business benefits of studying the communities, many of them are organised around a brand and product.

The virtual communities can be spontaneously formed or organised by the company. An example of the latter is the community created inside the Swatch's site (Bartoccini & Di Fraia, 2004, 196). The researcher could even create an ad hoc community, without rely on extant ones (spontaneous or organised), that would fit the research objectives. In creating a community, the researcher should follow the same rules that keep alive a normal community, such as the organisation of rituals that foster the member's identity in the group and their attendance, allowing for the formation of roles among the users (Kim, 2000).

Some community can exert a real power on the product. The fans of the famous movie series *Star Wars* (Cova, 2003) pushed the producers towards changes in the screenplay, reducing the role played by a character not loved by the fans. Also, when no power is exerted on the firms' choices, analysing a community of consumption can give the firm useful

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

insights about the tastes of the market (Prandelli & Verona, 2001). A virtual environment can be leveraged for innovations too (Sawhney, Prandelli, & Verona, 2003).

All these applications and forms of the communities call for an examination of research methods, as literature has begun to approach in a systematic way (Jones, 1999), seeking for those methods that best fit with the particular features of virtual sociality. Independently from the method applied, according to Bartoccini and Di Fraia (2004, 200-201) the virtual community has pros and cons for research as listed in Table 1.

METHODS OF RESEARCH

Questionnaire Survey

According to Dillman (cit. in Cobanoglu, Warde, & Moreo, 2001), the most relevant innovation in survey methods were introduced in the 1940s by the random sampling techniques and telephone interviews in the 1970s. The Internet should be a third wave of innovation for survey.

A questionnaire survey administered through the Web, specifically by e-mail, seems to have a clear advantage in efficiency and costs. It is very easy to send a questionnaire to huge numbers of addresses. An alternative method, even easier in its administration, can be that of posting the questions inside a Web page, asking the visitors to fill in the questionnaire. Some Web sites are beginning to offer Internet space for online surveys. A detailed segmentation of the population can be reached thanks to search engines or users lists. Moreover the anonymity and the lack of any sensory clues may push the respondent towards a more sincere and open attitude, reducing the incidence of socially desirable answers. The lack of a human interviewer limits also the mode effect, avoiding that the style and modality of the interviewers would affect the answers (Sparrow & Curtice, 2004). Another benefit is the asynchronous feature of the e-mail, allowing the respondent to answer the questions at her convenience and with calm reasoning.

Barry (2001) cites his expected difficulty in finding enough subjects for one of his Internet studies about ethnic minorities. He planned to study the acculturation of Arabic immigrants in the U.S. His study was conducted just after the Oklahoma bombing in 1995 that initially "resulted in widely publicized and unfounded speculation about the possible involvement of Middle Eastern terrorists" (Barry, 2001, 18). Due to that atmosphere some of the respondent expressed suspicion, even asking whether the researcher was affiliated whit some sort of police agency. Eventually, the anonymity of the Web-administered questionnaire allowed a very good response ratio and, above all, for a high quality of the answers provided. The respondents were quite sincere and deep in their answers. As Barry argues, "One potentially potent use of the Internet is that it facilitates self-exploration; it can serve as a safe vehicle for individuals to explore their identity. This is facilitated by a prevalent sense of anonymity, which often results in increased self disclosure and disinhibition" (Barry, 2001, 17).

Yet the quality of the results of an online questionnaire may be not high. Firstly, a sampling issue arises. The response rates of online questionnaires are lower than expected. Usually, people filter unsolicited mails, due to "spam" and viruses concerns. The fear of being cheated, even though the anonymity is assured, may be higher than in the off-line reality, since there is not an actual and reassuring interviewer. Moreover the Internet population is not representative of the entire population, but it is likely younger segment open to new technologies. The reliability of the answers received is not high as well. In fact, none can assure who actually answered the questions.

Table 1. Advantage and disadvantage of virtual communities for the research activity

Advantages	Disadvantages
High involvement by the members	Biased sample
Spontaneity of the information provided by the members	Fake identity
Archive of past exchanges	Overflow of material not tied to the research objective

Source: Adapted from Bartoccini and Di Fraia (2004, pp. 200-201)

Table 2. Comparison of survey by mail, fax, and web

	Mail	Fax	Web
Coverage	High	Low	Low
Speed	Low	Low	High
Return Cost	Preaddressed/Prestamped Evelope	Return Fax Number	No cost to the respondent
Incentives	Cash/Non cash incentives can be included	Coupons may be included	Coupons may be included
Wrong Addresses	Low	Low	High
Labour Needed	High	Medium	Low
Expertise to Construct	Low	Medium	High
Variable Cost for Each Survey*	About $1	About $0,50	No cost (US)

Riva, Teruzzi and Anolli (2003) compare questionnaires aimed at assessing psychological traits, administered through traditional paper-and-pencil form or through the Web. While the two samples differ, the validity of the results is not significantly different; anyway the authors suggest a particular care in assessing the validity of the online measurement and in sampling procedures. Cobanoglu, Warde and Moreo (2001) conduct a comparison among three survey methods: mail, fax, and Web. Table 2 is a synthesis of their features.

The coverage by a Web-based survey is considered low by the authors because many subjects that would belong to the population studied have no e-mail address or they change it (this frequently happens compared to common mail addresses). Comparing response speed, response rate, and costs of the three types of survey, the authors find that the quicker method in getting responses is fax, followed by Web and, as expected, normal mail. A Web survey collects the majority of responses in the first days. The Web is the best method if measured in response rate terms, followed by mail and fax. The same order holds for the costs, with Web-based survey being less expensive method. The findings of the three authors are useful in envisaging the respective pros and cons of different survey methods; yet, as mentioned by the authors themselves, the population chosen for their research (US hospitality educators) seriously bound the external validity of their findings. For instance, when addressed to a larger Internet population, the response rate of Web-based surveys receive quite a low response rate (Sparrow & Curtice; 2004).

Other researches warn against an indiscriminate use of Web-based polls if intended as a perfect substitute of telephone polls. Panels built online are not necessarily similar to those off-line, letting different results coming out on a range of issues. In a recent study, the authors "have found marked differences between the attitudes of those who respond to a conventional telephone poll and both those who say they would, and those who actually do, respond to an online poll" (Sparrow & Curtice, 2004, p. 40). Anyway, the e-polls can be quite predictive: the Harris Interactive survey conducted online few weeks before the 2000 U.S. presidential election was one of the most precise (Di Fraia, 2004).

Grandcolas, Rettie, and Marusenko (2003) point out four main source of error in a survey: coherence between the target population and the frame population, sampling, non-response, and measurement. They found that the main source of error is not related to the questionnaire administration mode, but to the sample bias. In fact, the sampling error in researches through Internet may be quite relevant and is usually the bigger flaw for e-research in general (Di Fraia, 2004).

In a virtual community the coverage error should be less relevant if the community is the object of research. In this case, the whole population is the community itself. Yet the response rate should be not high. A community usually has a sort of fence that none can trespass with unsolicited messages. As an example, a questionnaire aimed at measuring reciprocal trust of the member of a support newsgroup was administered by the writing author. The questionnaires returned were four out of the about 50 sent, a mere 8 percent; more importantly, the questionnaire returned were from the most disgruntled members of the community that saw the questionnaire as an opportunity to express their anger, rather than answering to the questions. This failure, both in quantity and quality, was due to the fact that the researcher did not participate to the group's life

before the questionnaire's submission. I was a stranger. This resistance by the community's members towards strangers may be particularly intense for groups, like that previously described, that deals with intimate and delicate topics like illnesses.

Experiment

Many of the benefits mentioned for questionnaires can be found for experiments, but this holds for limitations too. Modern experiments are often administered through computer interfaces. The Web, by this way, has an advantage, being already a computer-based context. But experiments must be conducted in a highly controlled environment, in order to get valid results. This cannot be achieved in a far-fetched contact like that assured by the Web. The experimenter cannot control possible factors that would interfere with the research. Moreover, "interactions between the construct being measured and the characteristics of the testing medium" (Riva, Teruzzi, & Anolli, 2003, p. 78) can occur, infringing the experiment's validity. It might be necessary to adjust the test to suit the Web's features.

Content Analysis

Multimedia technologies are quickly developing. Today the user can download music files, images, clips and even connect to live TV broadcast. Visual-based communities are growing. Still, the Internet is eminently a textual medium. The textual nature of Internet is true for the most part of virtual communities too. Some communities are simulations of the reality and the participants can depict themselves as personages (avatars). But the most part of the virtual communities are text-based, like the newsgroups. This feature fits with the content analysis method. Content analysis can be defined as "a research technique for the objective, systematic and quantitative description of the manifest content of communications" (Berelson, cited in Remenyi, 1992, p. 76; see also: Bailey, 1994; Berger 2000; Gunter 2000; Kassarjian, 1977). Really, content analysis can be fruitfully applied to non verbal content as well (like photographs) but it originates from textual studies and its elements (words, sentences, themes). The verbal expressions can have various forms: speeches, conversations, written texts (Schnurr, Rosenberg, & Oxam, 1992).

The literature has employed content analysis to study Web sites. For instance, content analysis has been applied to direct-to-consumer drug Web sites to assess their implication on public policy (Macias & Stavchansky Lewis, 2004), to hotels' Web sites to check their private policies (O'Connor, 2003), and to Internet advertisements (Luk, Chan, & Li, 2002). Narrative analysis has been used for online storytelling (Lee & Leets, 2002). The analysis of the content created inside virtual communities is less developed.

Content analysis can be divided into paradigmatic and sintagmatic analysis. In the paradigmatic approach, the meaning is built along the text, with addition of new elements. The meaning of the text is not in its element individually taken, but in the linear connection among them, in the development of the text. This approach is drawn from the narrative analysis of tales and other texts that build the sense through different episodes and personages. The sintagmatic approach extracts the meaning classifying the elements of the text independently from their position within it. The content of a virtual community is not in its single elements, but it is built through the interaction. The texts produced in virtual environment are not stand-alone posts; they form a net of questions and answers, statements and reactions, even flamed exchanges. The observer cannot validly catch the meaning without following the entire chain of exchanges. Along this path, content analysis in virtual communities is near discourse analysis, a branch of content and rhetorical analysis that further considers utterances as anchored to the contingent story of that specific exchange of communicative deeds. The researcher may not catch the real meaning of what is happening in a virtual community just extracting and classifying single words. She should follow the story of exchanges, pinpointing themes and personages as they develop. In this sense, content analysis should be integrated with an actual understanding of who is posting and which is the stage of the conversation. This deep understanding can be reached through the integration of content analysis with other methods, like netnography (see below).

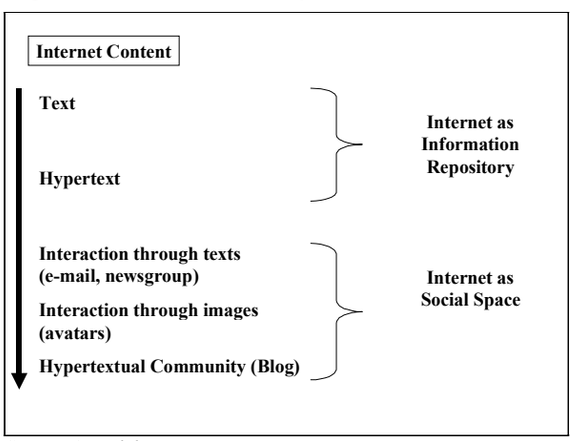
The difference between sintagmatic and paradigmatic approach is defied by hypertexts, a form of writing that is peculiar to the Internet. While interactivity, availability, non-intrusiveness, equality, and other features of Internet are known (Berthon, Pitt, Ewing, Jayaratna, & Ramaseshan, 2001), per-

haps hypertextuality is the main feature of the Web. Web sites are linked together and a single Web site in itself is structured as a hypertext.

In a hypertext, there is no linearity in the construction of meaning; indeed a meaning does not exist until the reader builds one with the act of reading. The traditional narrative analysis does not apply. The relevance of hypertextual communication is growing. Hypertextual structures are not limited to single text or to Web site links anymore. Blogs are a sort of hypertextual community, where personal Web sites are tied together through links. In such a community the meaning is not in a single text, word per Web site, but in the complex net of references and links. Traditional content analysis should develop new mean to study such a phenomenon, where the community is spread in many contributions in different locations of the Web space.

Thanks to the advancement in computer-mediated communication technology, new forms of virtual communities are springing, where the text is less relevant. They are graphic virtual realities (Kolko, 1999): visual-based communities where the textual element is less relevant. These communities are populated by avatars, visual representations of the users that interact among them and with the virtual environment. Content analysis, intended as textual tool, meets its boundaries. In this case the object of study is not the text, rather the choices made by the user about his/her avatar's features and any movement of the avatar. The design of the avatar can be considered a rhetorical act that should be studied by rhetoric.

Figure 1. The evolution of Internet content



Source: Our Elaboration

We can sketch the evolution of Internet content in Figure 1.

Netnography

A common thread of the methods illustrated above is the necessity for the researcher of being embedded in the community's life as a regular member. In fact, a questionnaire pushed forward by an unknown researcher may raise suspect and opposition within the community. The meaning of the exchange that occurs on the screen can be understood only by an actual member that knows the particular language employed in that community and the personages that live there. Kozinets (1998, 2002) has developed and applied the method called "netnography": ethnography applied to virtual communities (see also Brown, Kozinets, & Sherry, 2003, for an application). The researcher "lives" inside the virtual community, immersed in it, observing the dynamics, like an ethnographer lives in and observes a tribe or a social group. Through netnography the researcher gain a "thick" knowledge (Geertz, 1977) of the community. This "locality" of knowledge and the researcher's unavoidable subjectivity may impair the external validity of the netnographic study, yet offering valuable insights and knowledge. Kozinets (2002) provides suggestions about the steps that should be followed to have a good netnographic study: define a clear research question, choose a virtual group that suits the research, gain familiarity with the group, gather the data, and interpret them in a trustworthy manner.

As noticed before, the most difficult part in doing netnography is to gain a legitimate entrance into the community. The researcher must be very familiar with the participants of the community and the rules—mostly implicit—that apply (Kozinets, 2002, p. 5). Ethics ought to be a main focus for researcher; this is even more relevant for netnography, since this method raises new ethical issues: is the written material found in a newsgroup public? Which are the boundaries of the "informed consent" concept? (Kozinets, 2002, p. 9). The postings may be conceived by the virtual community members as exclusively addressed to the other participants, even though the Internet is an essentially public space. While this expectation can be irrational, the netnographer should respect it and she should address this facet, given that the "potential for 'netnography' to do harm is real

risk. For instance, if a marketing researcher were to publish sensitive information overheard in a chat room, this may lead to embarrassment or ostracism if an associated person's identity was discerned" (Kozinets, 2002, p. 9).

CONCLUSION

The Internet and virtual communities offer a very wide space of research. Texts can be downloaded, people can be contacted, and sites can be thoroughly analysed. For some type of research, the Internet is really a frictionless space where data are not a troubling part of the research. Yet this easiness can be the path towards badly planned research, since the lots of data may push the researcher towards less than careful attention to the method. Therefore the focus on research methods on the Internet is a central issue.

Virtual communities are the growing part of the Internet's development. What is the best research method to study a virtual community? From the argumentation shown above, every method has pros and cons. The following table synthesises some features of the main research methods applied to the Internet.

The method that seems to emerge for the future as particularly fitting with the virtual communities realm is netnography. Anyway the researcher should choose the method that fits with his particular research propositions. A rigorous execution of the method is also at the centre of a valid and reliable research.

A case of research may be helpful in showing the difficulties of online studies. The writing author was interested in measuring trust inside virtual communities. Trust among members is in fact a basic requirement to build a long-lasting and efficient community. The virtual community chosen was a support group for subjects with a particular disease. Groups that deal with disease strongly need trust, even though the disease is neither very serious nor too personal to speak about. The subject should have trust in the fact that others will not exploit or deride that delicate and personal opening; he should trust the medical solutions suggested by unknown persons.

A trust scale to submit to the virtual group under the form of questionnaire was unsuccessful, as shown before, since the response rate was quite low and the answers biased (most people answered just to vent anger towards some group's member).

An experiment would have been coherent with the game theory approach that trust studies can take. Yet, the lack of any control on the subjects led to discard this method.

As to content analysis, trust is a construct that, paradoxically, is present when no one speaks about it: betray of trust would elicit strong reactions and "flaming", while trust would not be explicitly stated by a subject. The mere fact that an individual speaks about her disease is a sign of trust towards the others, but this does not suffice for an exact measurement of the construct. Moreover, defying any e-research method, the real and deep trust would occur when subjects find each other so trustworthy that they meet off line, continuing there their interaction. In this case, the content analysis would totally miss its material of study. Finally, netnography seemed to be the best way to achieve conclusions about trust and its drivers in the community. But the difficulty in doing netnographic research was in the immersion in the group's experience. The nuances of meanings, the implicit codes of language, the bundle of emotions of persons that often

Table 3. Features of different method of research applied to a virtual community

	Advantages	Disadvantages
Survey/Questionnaire	- Easiness in administering the questionnaire	- Difficult 'entrée' into the virtual community - Low response rate - Lack of control on who actually answer - Biased sample
Experiment	- Computer-based format suitable for Internet	- Difficult 'entrée' into the virtual community - Lack of control on who actually is the subject - Lack of environmental control
Content Analysis	- Most virtual communities are text-based - Unobtrusive - Lots of data available	- The exchanges are discourses, rather than words
Netnography	- High level of understanding of the community by the researcher	- Local knowledge, low external validity - Risk of subjectivity

for their first time revealed their disease: all this is truly understood only by a subject with that problem, putting a sort of unavoidable distance between the researcher and the group.

The case briefly outlined shows the issues that such a new environment like virtual communities raise for the research.

REFERENCES

Alba, J., Lynch, J., Weitz, B., Janiszevski, C., Lutz, R., Sawyer, A., & Wood, S. (1997). Interactive home shopping: Consumer, retailer, and manufacturer incentives to participate in electronic marketplaces, *Journal of Marketing, 61*(July), 38-53.

Bailey, K.D. (1994). *Methods of social research*. MacMillan.

Barry, D.T. (2001). Assessing culture via the Internet: Methods and techniques for psychological research. *CyberPsychology & Behavior, 4*(1), 17-21.

Bartoccini, E. & Di Fraia, G. (2004). Le Comunità Virtuali come Ambienti di Rilevazione. In G. Di Fraia (Ed.), *E-research: Internet per la Ricerca Sociale e di Mercato*. Editori Laterza, 188-201.

Berger, A.A. (2000). *Media and communication research methods*. SAGE.

Berthon, P., Pitt, L., Ewing, M., Jayaratna, N., & Ramaseshan, B. (2001). Positioning in cyberspace: Evaluating telecom Web sites using correspondence analysis. In O. Lee (Ed.), *Internet marketing research: Theory and practice*. Hershey, PA: Idea Group Publishing, 77-92.

Brown, S., Kozinets, R.V., & Sherry, J.F. (2003). Sell me the old, old story: Retromarketing management and the art of brand revival. *Journal of Consumer Behaviour, 2*, 133-147.

Castells, M. (1996). *The rise of the network society*. Blackwell Publishers.

Cobanoglu, C., Warde, B., & Moreo, P.J. (2001). A comparison of mail, fax and Web-based survey methods. *International Journal of Market Research, 43*(4), 441-452.

Cova, B. (2003). *Il marketing tribale*. Il Sole 24 Ore.

Di Fraia, G. (2004). Validità e attendibilità delle ricerche online. In G. Di Fraia (Ed.), *E-research. Internet per la ricerca sociale e di mercato*. Editori Laterza, 33-51.

Donath, J.S. (1999). Identity and deception in the virtual community. In M. Smith & P. Kollock (Eds.), *Communities in cyberspace*. London: Routledge, 29-59.

Geertz, C. (1977). *Interpretation of cultures*. Basic Books.

Grandcolas, U., Rettie, R., & Marusenko, K. (2003). Web survey bias: Sample or mode effect? *Journal of Marketing Management, 19*, 541-561.

Gunter, B. (2000). *Media research methods*. SAGE Publications.

Hoffman, D.L. & Novak, T.P. (1996). Marketing in hypermedia computer-mediated environment. *Journal of Marketing, 60*(July), 50-68.

Jones, S. (1999). *Doing Internet research: Critical issues and methods for examining the Net*. Thousand Oaks, CA: Sage.

Kassarjian, H.H. (1977). Content analysis in consumer research. *Journal of Consumer Research, 4*, 8-18.

Kim, A.J. (2000). *Costruire comunità Web*. Apogeo.

Kolko, B.E. (1999). Representing bodies in virtual space: The rhetoric of avatar design. *The Information Society, 15*(3), 177-186.

Kozinets, R.V. (1998). On netnography: Initial reflections on consumer research investigations of cyberculture. In J. Alba & W. Hutchinson (Eds.), *Advances in consumer research, Volume 25*. Provo, UT: Association for Consumer Research, 366-371.

Kozinets, R.V. (2002). The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research, 39*(February), 61-72.

Lee, E. & Leets L. (2002). Persuasive storytelling by hate groups online. *American Behavioral Scientist, 45*(6), 927-957.

Luk, S.T.K., Chan, W.P.S., & Li, E.L.Y. (2002). The content of Internet advertisements and its impact on awareness and selling performance. *Journal of Marketing Management, 18*, 693-719.

Macias, W. & Stavchansky, L.L. (2004). A content analysis of direct-to-consumer (DTC) prescription drug Web sites. *Journal of Advertising, 32*(4), 43-56.

O'Connor, P. (2003). What happens to my information if I make a hotel booking online: An analysis of online privacy policy use, content and compliance by the International Hotels Company. *Journal of Service Research, 3*(2), 5-28.

Prandelli, E. & Verona, G. (2001). *Marketing in rete: Analisi e decisioni nell'economia digitale*. McGraw-Hill.

Rayport, J.F. & Sviokla, J.J. (1994). Managing in the marketspace. *Harvard Business Review, November/December*, 141-150.

Remenyi, D. (1992). Researching information systems: Data analysis methodology using content and correspondence analysis. *Journal of Information Technology, 7*, 76-86.

Rheingold, H. (2000). *The virtual community: Homesteading on the electronic frontier, Revised edition*. MIT Press.

Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of Internet in psychological research: Comparison of online and offline questionnaires. *CyberPsychology and Behavior, 6*(1), 73-79.

Roversi, A. (2001). *Chat line*. Il Mulino.

Sawhney, M., Prandelli, E., & Verona, G. (2003). The power of innomediation. *Sloan Management Review, 44*(2), 77-82.

Schnurr, P.P., Rosenberg, S.D., & Oxam, T.E. (1992). Comparison of TAT and free speech techniques for eliciting source material in computerized content analysis. *Journal of Personality Assessment, 58*(2), 311-325.

Smith, M. (1999). Invisible crowds in cyberspace: Mapping the social structure of the Usenet. In M. Smith, M. & P. Kollock (Eds.), *Communities in cyberspace*. London: Routledge.

Sparrow, N. & Curtice, J. (2004). Measuring the attitudes of the general public via Internet polls: An evaluation. *International Journal of Market Research, 46*(1), 23-44.

Stafford, T.F. & Stafford M.R. (2001). Investigating social motivations for Internet use. In O. Lee (Ed.), *Internet marketing research: Theory and practice*. Hershey, PA: Idea Group Publishing.

Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet*. Simon & Schuster.

KEY TERMS

Avatar: Personification of a user in a graphic virtual reality. An avatar can be an icon, an image, or a character, and it interacts with other avatars in the shared virtual reality. The term is drawn from the Hindu culture, where it refers to the incarnation of a deity.

Content Analysis: Objective, systematic, and quantitative analysis of communication content. The unit of measure can be the single words, sentences, or themes. In order to raise the reliability, two or more coders should apply

Experiment: Research method in which the researcher manipulates some independent variables to measure the effects on a dependent variable.

MUD: Multi-User Dungeon; a virtual space where subjects play a game similar to an arcade, interacting through textual and visual tools. In a MUD it is usual to experience a hierarchy.

Netnography: Method of online research developed by Robert Kozinets (Kellogg Graduate School of Business, Northwestern University, Evanston). It consists in ethnography adapted to the study of online communities. The researcher assumes the role of a regular member of the community (but, for ethical reasons, she/he should disclose her/his role).

Survey: Measurement procedure under the form of questions asked by respondents. The questions can be addressed through a written questionnaire that the respondent has to fill in or through a personal interview (following or not a written guideline). The items of the questionnaire can be open or multiple choice.

Migration to IP Telephony

Khaled A. Shuaib
United Arab Emirates University, UAE

INTRODUCTION

There are mainly two types of used communication systems; circuit switched and packet switched networks. In circuit switched networks, there must be a dedicated path and a sequence of connected links between the calling and called stations. A connection with the proper resources has to be established prior to the start of information exchange. An example of circuit switched network is the phone network. On the other hand, packet switched networks rely on allowing multiple communicating end systems to share the entire or part of a path simultaneously. The Internet, a world wide computer network, is based on the concept of packet switching empowered by the Internet Protocol (IP). IP is basically a transmission mechanism used by devices communicating in a network as part of a protocol suite.

IP telephony is a technology based on the integration of telephony and other services with packet switched data network. IP telephony utilizes packet switched networks and implies multimedia (voice, video, and data) communication over IP or as it is often called converged services [Ibe, 2001; IP Telephony Group of Experts, 2001] allowing simultaneous communication between devices such as computers or IP phones. IP telephony becomes a very popular concept and an important technology that defines communication between individuals and organizations, public, and private [Gillett, Lehr, & Osorio, 2000]. Converged voice, video, and data IP based telephony is considered to be relatively new with respect to circuit switched telephone systems; however it is already being recognized as one of the current revolutionary technologies of the 21st century.

Several public and private institutions in different countries are considering to migrate their telephone systems from legacy circuit switching to packet switching using IP telephony. In most cases, a public institution has two separate networks: data and voice. The voice network can be depicted as one central office with many other branch offices scattered over one or more states, countries, or cities. Typically, these branch offices within a limited geographical region are connected to each other and to the central office via limited bandwidth leased telephone lines provided by the Public Switched Telephone Network (PSTN) for a particular cost. In some cases, where the branch offices are scattered beyond the geographical limits allowed, there is a complete disconnect between these branches and phone communications will be made at the cost of long distance calls.

In general, most organizations have a LAN for data traffic within the central office which is extended via the local carrier company data network to other branch offices over leased lines or Permanent Virtual Circuits (PVCs) composing the MAN of the organization. This network is only used for the transport of data, with no voice traffic transferred over it regardless of its capabilities. Figure 1 shows a typical current networking infrastructure of an institution with multiple sites in different regions.

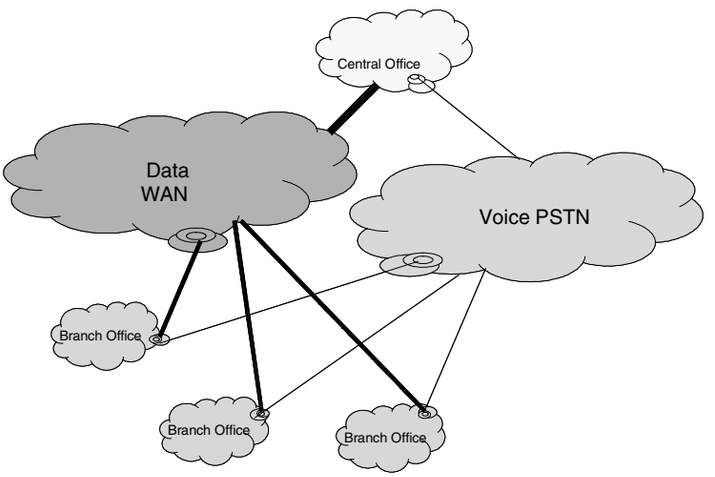
The goal of any institution is to integrate voice, video, and data over a single network infrastructure while maintaining quality, reliability, and affordability. Converged voice, video, and data Internet protocol based telephone systems are currently being thought of and accepted as the next generation platform replacing legacy PBX telephone system which has been providing us with great and reliable services over the last decades. A well built and designed LAN/WAN infrastructure is the key to providing acceptable, scalable, reliable, and affordable IP telephony (Keagy, 2000; Lacava, 2002).

MIGRATION PATHS

Today there exist two trends to replace the legacy PBX system and migrate to IP telephony: Converged IP-PBX and IP-PBX. The IP-PBX system can be described as a voice communication system that supports IP telephony operations and functions using fully integrated system design elements, both hard-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. A typical network for a multi site institution



ware and software utilizing a LAN/WAN infrastructure of an organization (Insight Research Corporation, 2003). This type of a system or solution is being mostly favored by packet switched or data network equipment makers.

The Converged IP-PBX system is based on a circuit-switched network design, but can be equipped with fully integrated media gateway port interface circuit cards to support IP stations and IP trunk ports. The Converged IP-PBX is best described as a bridge between the legacy PBX system and the IP-PBX system. This type of a system/solution is being favored by circuit switched equipment (legacy PBX) makers.

Each of these systems carries its own advantages and disadvantages which will vary based on the proposed implementation and the use of the system (Considerations IDC Executive Brief, 2002). Both systems are being offered and sold by many telecom vendors like Cisco, Nortel, Lucent, Avaya, Alcatel, Siemans, and so on; however, the migration path to any of the two new systems is a unique process that is dependent on many factors (Yankee, Group, 2003). Accordingly, to switch from a traditional circuit switching infrastructure to IP telephony, there are two converging paths, IP-PBX and Converged IP-PBX; however, a mixed solution that can utilize the best of both paths is possible (Cisco systems, 2000; Lucent technology, 2003; Nortel Networks white paper, 2003; Thurston, Hall, & Kwiatkowski, 2002).

A chosen solution, that will provide converged services over IP, must scale to PSTN call volumes, offer PSTN call quality, reliability, and equivalent services. It must also support new and innovative significant other optional services. The choice of any solution is usually coupled with cost justification, not just based on the initial cost of investment, but also based on the long term savings on capital cost, operations, and maintenance as well as other realized factors such as work productivity, work time saving, travel expenses, employee retention, and so on (Cisco Systems, 2001; O'Malley, 2003).

IP-PBX: FEASIBILITY, ADVANTAGES AND RESERVATIONS

The IP-PBX solution fully utilizes a packet switched network for the deployment of integrated services over a private enterprise WAN network (Christensen, 2001).

Feasibility

This solution can apply best for some situations such as:

1. A green field investment, where a new building is being considered.
2. Building a new or upgrading the infrastructure of a data network including LANs, due to the exponential growth of data network.

The above two points can be justified based on the following reasons:

1. New wiring and equipment is needed.
2. There might be minimal used equipment that can be leveraged.
3. The saving of wiring for circuit switching.
4. The elimination of any needs for two sets of staffs for circuit and packet switched networks.
5. The willingness of employees to start with a new system.
6. The desire to use the most advanced and futuristic technology available.
7. If the cost of the alternative solution (Converged IP-PBX) is higher than the IP-PBX solution, which might be true for certain systems and under certain configurations.

Advantages

A typical network topology utilizing the IP-PBX system is depicted in Figure 2. Such a system will provide an organization with many benefits, such as:

1. The use of an IP-PBX system will mean the use of one single infrastructure and cable for voice, video, and data
2. Utilization of the costly bandwidth available on the LAN/WAN data network
3. The reduction of operation and maintenance cost of the network especially in a multi-site organization, since operations can be centralized and managed from a single site
4. Easy and less costly to upgrade over the life of the product
5. Full-pledged new features that can ease the management of employees and tasks
6. Flexible accessibility of the system at any time from any networked place
7. Reduce travel time for cross sites meetings due the use of efficient and scalable voice and video conferencing capabilities.
8. Flexibility in employee re-allocation within and across connected sites (play and plug telephony)
9. Integration or unification of message systems (voice, e-mail, and fax)
10. Increases knowledge sharing across the organization through the ease of using video conferencing and database sharing

Based on the chosen vendor, integrating analog and digital phones for an additional cost of voice gateway devices as shown in Figure 3 is possible.

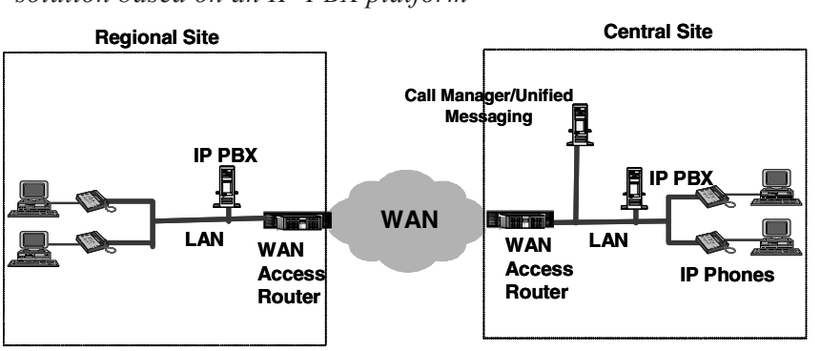
This type of solution will utilize the WAN infrastructure for the transport of data, voice, and video; however, any calls to the outside of the WAN will have to be routed through the PSTN for an addition cost.

Reservations

The previous benefits can only be fully realized when certain conditions and measures are considered and not compromised. These measures can be summarized as follows.

1. A well designed and voice-capable LAN that implements QoS disciplines to prioritize voice traffic

Figure 2. An all IP solution based on an IP-PBX platform



2. An emergency and very limited minimal implementation of a legacy PBX system that is only used by certain individuals in the case of any data network failure
3. Redundancy of LAN to WAN access routers/switches and call servers, which can add to the total implementation cost
4. Strict Service Level Agreement (SLA) with the local ISP to guarantee 99.999 percent availability (Cisco Systems, 2002).
5. A reliable WAN configuration with build in redundancy and QoS support
6. Sufficient LAN and WAN resources, that is, network capacity to handle worse case scenarios and future growth
7. Trained staff for network management and maintenance
8. Compatibility among the devices of all sites
9. Security must not be compromised, and every measure must be taken to protect the network. An IP-PBX network is probably as secure as your office PC today. If the network is infected with a virus, there might be a total network melt down.
10. Employees trained and ready to use the features provided by such a solution
11. Willingness of users to tolerate possible worse voice quality during certain times
12. Willingness of the organization to tolerate possible network down time upon an unexpected network failure

It must be noted that in most cases, it would be hard to balance the cost and benefits of such a network while having a mixed IP-PBX/Legacy network.

CONVERGED IP-PBX: FEASIBILITY, ADVANATGES AND RESERVATIONS

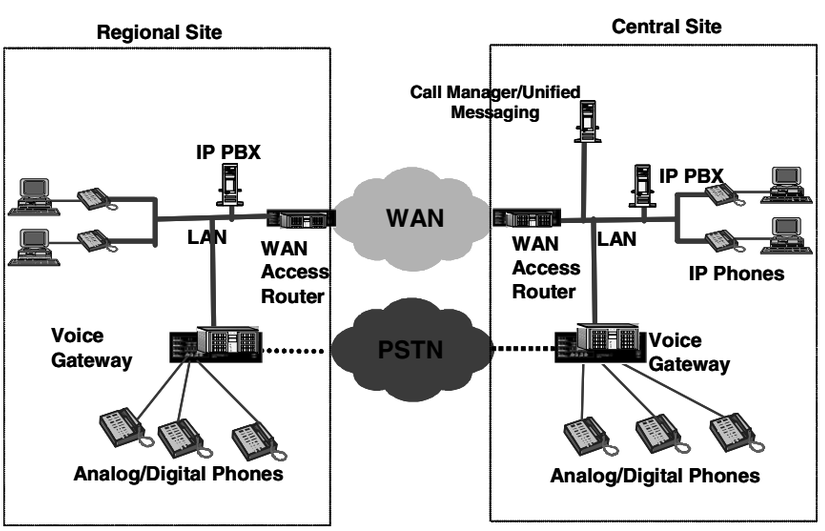
The Converged IP-PBX system can be looked at as a hybrid circuit switched/packet switched system that is often used for the gradual upgrading and shifting from a circuit switched environment to a packet switched one by either upgrading existing legacy PBX equipment to handle IP-based voice traffic or by installing new Converged IP-PBX equipment that can handle both legacy and IP voice traffic (Nortel Networks, 2001; Sulkin, 2001).

Feasability

This solution is usually attractive to organizations when one or more of the following is true:

1. The organization is interested in maintaining investments in legacy PBX equipment.
2. The organization is not ready to completely migrate to a total IP packet switched solution.
3. The cost of integrating or installing Converged IP-PBX is much less than that of an IP-PBX solution.

Figure 3. Integrating analog and digital phones in IP-PBX solution using voice gateways



4. The organization has no plans to upgrade LAN/WAN infrastructure which is currently only capable of supporting limited voice applications.
5. Compatibility issues exist between devices at the different sites of the organization.
6. The willingness of the organization to maintain skilled operation and maintenance staffing in multiple sites for both legacy and new converged systems.

Advantages

Most of the benefits realized by the IP-PBX solution can also be realized with a Converged IP-PBX with some limited to the IP telephony users only. These benefits can be summarized as follows:

1. The use of a Converged IP-PBX solution can be realized, with limitations, to use one single infrastructure for voice, video, and data.
2. Utilization of the costly bandwidth available on the LAN/WAN data network
3. Full pledged new features that can ease the management of employees and tasks (utilized by IP telephony users)
4. Flexible accessibility of the system at any time from any networked place (utilized by IP-phone users)
5. Reduce travel time for cross sites meetings due the use of voice and video conferencing capabilities (utilized by IP telephony users)
6. Flexibility in employee re-allocation within and across connected sites (utilized by IP telephony users)
7. Integration or unification of message systems (voice, e-mail, and fax) (utilized by IP telephony users)
8. Increased reliability, provided that the Converged IP-PBX system is capable of utilizing the PSTN incase of a data network failure event
9. Allows for the gradual migration from legacy PBX to IP telephony

Reservations

Similar to the IP-PBX solution, the mentioned benefits can only be fully realized when certain conditions, reservations and measures are considered and not compromised. These measures are the same as of those mentioned for the IP-PBX with the following differences or additions:

1. Skilled staffs for both legacy PBX and IP telephony network operation and management. For the IP-PBX, no legacy PBX staff is needed.
2. Willingness of the IP telephony users to tolerate possible worse voice quality during certain times. In the IP PBX all users will be IP telephony users.
3. Upgradeable Legacy PBXs to Converged IP-PBXs to leverage previous investment

Figure 4. A Converged IP-PBX solution with no IP phones

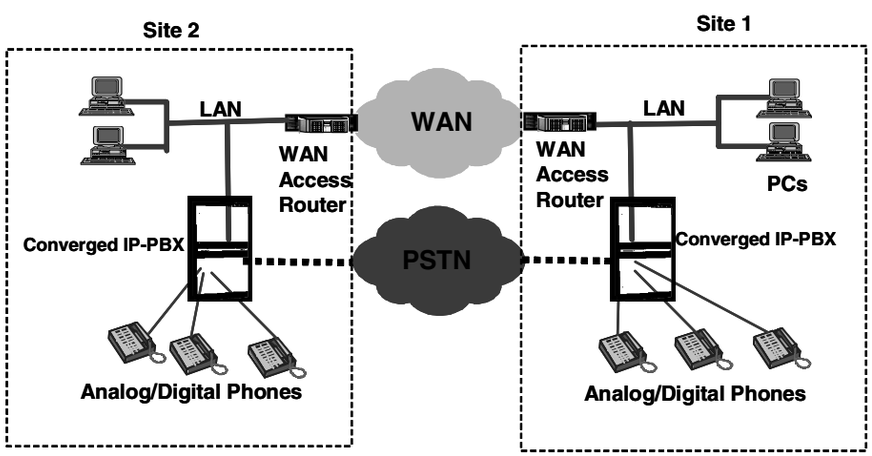
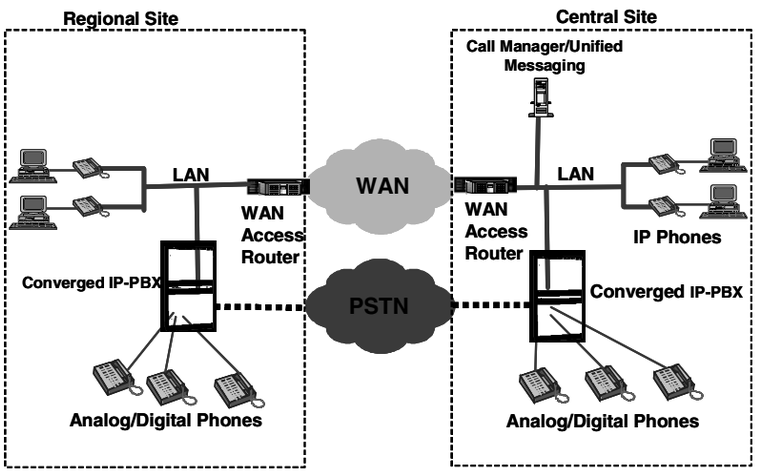


Figure 5. A Converged IP-PBX solution with IP phones



4. Converged IP-PBX compatibility among the devices at the different sites of the organization
5. Use of scalable converged IP-PBX systems that can handle expected growth of the use of IP-phones in the organization while maintaining the cost benefits over an IP-PBX system

This type of solution will have utilization of the WAN infrastructure for the transport of data, voice, and video. The use of the PSTN network will be reduced but not totally eliminated and that will depends on the number of converged users, the design and capacity of the converged network and the call volume to sites which are not voice over IP ready due to LAN or WAN issues. Figure 4 show a typical network configured with Converged IP-PBX equipment to carry traffic over a WAN. In this figure, no IP phones are being used, however the Converged IP-PBX device must be equipped with voice gateway cards for IP trunk circuit connections for non-IP stations. In Figure 5, IP phones are added along with needed call management equipment and software.

For a Converged IP-PBX to be able to handle IP phones, additional voice gateway cards are needed. The cost of these voice gateway cards could add up significantly if the number of IP phone users increases. The economic feasibility of a Converged IP-PBX network will be realized as long as current investment in legacy PBX equipment can be leveraged, the use of IP phone users is limited (to limit the cost of voice gateways) and there is no or minimal need to upgrade a LAN/WAN infrastructure which is voice over IP capable.

CONCLUSION

The IP-PBX solution is described as a voice communication system that supports IP telephony operations and functions using fully integrated system design elements, both hardware and software utilizing a LAN/WAN infrastructure. The Converged IP-PBX solution is based on a circuit-switched network design, but can be equipped with fully integrated media gateway port interfaces to support IP stations and IP trunk ports.

The Converged IP-PBX is best described as a bridge between the legacy PBX system and the IP-PBX system. The migration path from legacy telephone systems to new voice over IP capable systems is unique per organization and should be designed while considering all aspects involved such as: initial and long term cost, features, benefits, risks, feasibility, and future growth. In general, we suggest one of the following three scenarios

First Scenario

Assuming that:

- The initial use of IP-phone will be limited or not desired;

- All or the majority of regional offices has already invested in legacy PBX systems which are upgradeable to Converged IP-PBXs;
- There is no interoperability issues integrating these offices via the converged IP-PBXs;
- The current LAN infrastructure at the majority of regional offices are or can be upgraded with minimal cost to carry voice over IP traffic;
- There exist a capable WAN infrastructure; and
- There is no interoperability issues integrating upgradeable legacy PBX to Converged IP-PBX systems, with new Converged IP-PBX systems to be installed in new buildings, then our recommendation would be to implement a Converged IP-PBX solution while emphasizing the reservations in this document with that regard.

Second Scenario

Assuming that:

- The initial use of IP phones is favored and very limited analog and digital phones are required;
- The main or central site of the organization is moving into a new building with no communication equipment or wiring;
- More of the regional offices are also being moved into new buildings with no communication equipments and wiring;
- The majority of LANs in the regional offices need to be upgraded to support growth in data traffic and to carry converged IP traffic;
- There exists a capable WAN infrastructure or a new one is proposed to be built; and
- Communications can be maintained between sites during the migration period, then our recommendation would be to implement an IP-PBX solution while emphasizing the reservations listed in this document with that regard.

Third Scenario

Assuming that:

- The initial use of IP phones is favored in new or totally renovated sites;
- The main or central site of the organization is moving into a new building with no communication equipment or wiring;
- The current LAN infrastructure at the majority of regional offices are or can be upgraded with minimal cost to carry converged IP traffic;
- All or the majority of regional offices has already invested in legacy PBX systems;
- Interoperability issues between an upgraded legacy PBX system to a Converged IP-PBX system, a new Converged IP-PBX system and an IP-PBX system can be overcome (this could be minimized by choosing a vendor that can provide all needed platforms or two vendors who can seamlessly interoperate); and
- There exist a capable WAN infrastructure, then our recommendation would be to implement Converged IP-PBX systems in old regional offices and IP-PBX systems in new buildings while emphasizing the reservations listed in this document with regard to both technologies.

REFERENCES

Christensen, S. (2001). Voice-over IP Solutions. *Juniper Networks white paper.*

Cisco Systems white paper (2000). VoIP/VoFR aggregation and tandem BX bypass on the Cisco 7200 and 7500.

Cisco Systems white paper (2001). The strategic and financial justifications for convergence.

Cisco Systems white paper (2002). IP telephony: The five nines story.

Considerations IDC Executive Brief (2002) Is IP telephony right for me? Network choices and customer. IDC.

Gillett, S., Lehr, W., & Osorio, C. (2000). Local government broadband initiatives. Presented at TPRC, Alexandria, VA.

Ibe, O. (2001). *Converged network architectures: Delivering voice- and data-over IP, ATM, and frame relay.* Wiley.

Insight Research Corporation (2003). IP PBX and IP Centrex: Growth of VoIP in the enterprise 2004-2009, a market research report.

IP TELEPHONY Group of Experts (2001). Technical aspects, ITU, 3rd Experts Group Meeting on Opinion D Part 3 (ITU-D) Geneva.

Keagy, S. (2000). *Integrating voice and data networks.* Cisco Press.

Lacava, G. (2002). Voice over IP: An overview for enterprise oranizations and carriers. *INS white paper.*

Lucent technology white paper (2003). PBX versus IP PBX.

Nortel Networks white paper (2001). Voice over IP Solutions for Enterprise.

Nortel Networks white paper (2003). Circuit to Packet Evolution.

O'malley, S. (2003). *Enterprise IP telephony: Evaluating the options.* Internet Telephony.

Sulkin, A. (2001). *Manageable migration: The IP - enabled PBX system.* TEQConsult Group.

Thurston A., Hall P., & Kwiatkowski, A. (2002). Enterprise IP voice: Strategies for service providers. *OVUM, white paper.*

Yankee, Group (2003). The PBX alternative: Hosted IP telephony.

KEY TERMS

IP: Internet or Internetworking Protocol. A set of rules that defines how transmission of data is carried over a packet-switched network.

ISP: Internet Service Provider; usually a company that provides users with Internet access.

LAN: Local Area Network, refers to a network connecting devices inside a single building or inside building close to each other.

MAN: Metropolitan Area Network; refers to a network of devices that is spanning the size of a city.

PBX: Private Branch Exchange; a private telephone network used within an enterprise.

PSTN: Public Switched Telephone Network; the well known classical telephone network.

PVC: Permanent Virtual Circuit; a virtual connection between two communicating devices on a network.

QoS: Quality of Service; refers to the quality of an application when transported over a network.

SLA: Service Level Agreement; an agreement between an Internet service provide and a customer regarding the type and quality of the provided services.

WAN: Wide Area Network; refers to a network that spans a large geographical area or distance.

Mobile Ad Hoc Network

Subhankar Dhar
San Jose State University, USA

INTRODUCTION

A mobile ad hoc network (MANET) is a temporary, self-organizing network of wireless mobile nodes without the support of any existing infrastructure that may be readily available on conventional networks. It allows various devices to form a network in areas where no communication infrastructure exists. Although there are many problems and challenges that need to be solved before the large-scale deployment of an MANET, small and medium-sized MANETs can be easily deployed.

The motivation and development of MANET was mainly triggered by Department of Defense (DoD)-sponsored research work for military applications (Freebersyser and Leiner, 2002). In addition, ad hoc applications for mobile and dynamic environments are also driving the growth of these networks (Illyas, 2003; Perkins, 2002; Toh, 2002). As the number of applications of wireless ad hoc networks grows, the size of the network varies greatly from a network of several mobile computers in a classroom to a network of hundreds of mobile units deployed in a battlefield, for example. The variability in the network size is also true for a particular network over the course of time; a network of a thousand nodes may be split into a number of smaller networks of a few hundred nodes or vice versa as the nodes dynamically move around a deployed area.

Ad hoc networks not only have the traditional problems of wireless communications like power management, security, and bandwidth optimization, but also the lack of any fixed infrastructure, and their multihop nature poses new research problems. For example, routing, topology maintenance, location management, and device discovery, to name a few, are important problems and are still active areas of research (Wu & Stojmenovic, 2004).

Characteristics of MANET

- Mobile: The nodes may not be static in space and time, resulting in a dynamic network topology.
- Wireless: MANET uses a wireless medium to transmit and receive data.
- Distributed: MANET has no centralized control.
- Self-organizing: It is self-organizing in nature.
-

A message from the source node to destination node goes through multiple nodes because of the limited transmission radius.

- Scarce resources: Bandwidth and energy are scarce resources.
- Temporary: MANET is temporary in nature.
- Rapidly deployable: MANET has no base station and, thus, is rapidly deployable.
- Neighborhood awareness: Host connections in MANET are based on geographical distance.

SOME BUSINESS AND COMMERCIAL APPLICATIONS OF MANET

An ad hoc application is a self-organizing application consisting of mobile devices forming a peer-to-peer network where communications are possible because of the proximity of the devices within a physical distance. MANET can be used to form the basic infrastructure for ad hoc applications.

Some typical applications are as follows:

- **Personal-area and home networking:** Ad hoc networks are quite suitable for home as well as personal-area networking (PAN) applications. Mobile devices with Bluetooth or WLAN (wireless local-area network) cards can be easily configured to form an ad hoc network. With

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Internet connectivity at home, these devices can easily be connected to the Internet. Hence, the use of these kinds of ad hoc networks has practical applications and usability.

- **Emergency services:** When the existing network infrastructure ceases to operate or is damaged due to some kind of disaster, ad hoc networks enables one to build a network and they provide solutions to emergency services.
- **Military applications:** On the battlefield, MANET can be deployed for communications among the soldiers in the field. Different military units are expected to communicate and cooperate with each other within a specified area. In these kinds of low-mobility environments, MANET is used for communications where virtually no network infrastructure is available. For example, a mesh network is an ad hoc peer-to-peer, multihop network with no infrastructure. The important features are its low cost, and nodes that are mobile, self-organized, self-balancing, and self-healing. It is easy to scale. A good example is SLICE (soldier-level integrated communications environment), a research project sponsored by DARPA (Defense Advanced Research Projects Agency) in this area for this need. The idea is that every soldier is equipped with a mobile PC (personal computer) with a headset and a microphone. SLICE is supposed to create mesh networks that handle voice communications while mapping whereabouts of soldiers and their companions.
- **Ubiquitous and embedded computing applications:** With the emergence of new generations of intelligent, portable mobile devices, ubiquitous computing is becoming a reality. As predicted by some researchers (Weiser, 1993), ubiquitous computers will be around us, always doing some tasks for us without our conscious effort. These machines will also react to changing environments and work accordingly. These mobile devices will form an ad hoc network and gather various localized information, sometimes informing the users automatically.
- **Location-based services:** MANET, when integrated with location-based information, provides useful services. GPS (Global Positioning System), a satellite-based radio navigation system, is a very effective tool to determine the physical location of a device. A mobile host in a MANET, when connected to a GPS receiver, will be able to determine its current physical location. A good example is that a group of tourists using PDAs (personal digital assistants) with wireless LAN cards installed in them along with GPS connectivity can form a MANET. These tourists can then exchange messages and locate each other using this MANET. Also, vehicles on a highway can form an ad hoc network to exchange traffic information.
- **Sensor network:** It is a special kind of hybrid ad hoc network. There is a growing number of practical applications of tiny sensors in various situations. These inexpensive devices, once deployed, can offer accurate information about temperature, detect chemicals and critical environment conditions (e.g., generate wild-fire alarms), monitor certain behavior patterns like the movements of some animals, and so forth. In addition, these devices can also be used for security applications. However, these sensors, once deployed, have limited battery power, and the lifetime of the battery may determine the sensor's lifetime. Recently, several government agencies (e.g., NSF [National Science Foundation]) have funded research projects on sensor networks.

MAC-LAYER PROTOCOLS FOR MANET

An ad hoc network can be implemented very easily using the IEEE 802.11 standard for WLAN. Since the mobile nodes in WLAN use a common transmission medium, the transmissions of the nodes have to be coordinated by the MAC (media-access control) protocol. Here we summarize the MAC-layer protocols.

- **Carrier-sense multiple access (CSMA):** Carrier-sense multiple-access protocols were proposed in the 1970s and have been used in a number of packet radio networks in the past. These protocols attempt to prevent a station from transmitting simultaneously with other stations within its transmitting range by requiring each station to listen to the channel before transmitting. Because of radio hardware char-

acteristics, a station cannot transmit and listen to the channel simultaneously. This is why more improved protocols such as CSMA/CD (collision detection) cannot be used in single-channel radio networks. However, CSMA performs reasonably well except in some circumstances where multiple stations that are within range of the same receivers cannot detect one another's transmissions. This problem is generally called a hidden-terminal problem, which degrades the performance of CSMA significantly as collision cannot be avoided, in this case, making the protocol behave like the pure ALOHA protocol (Fullmer & Garcia-Luna-Aceves, 1995).

- **Multiple access with collision avoidance (MACA):** In 1990, Phil Karn proposed MACA to address the hidden-terminal problem (Karn, 1992). Most hidden-node problems are solved by this approach and collisions are avoided.
- **Multiple access with collision avoidance for wireless LANs (MACAW):** A group of researchers, in 1994, proposed MACAW to improve the efficiency of MACA by adding a retransmission mechanism to the MAC layer (Bharghavan, Demers, Shenker, & Zhang, 1994).
- **Floor-acquisition multiple access (FAMA):** A general problem of MACA-based protocols was the collision of control packets at the beginning of each transmission as all terminals intending to transmit sends out RTS (request-to-transmit) signals. In 1995, another protocol called FAMA was proposed, which combined CSMA and MACA into one protocol where each terminal senses the channel for a given waiting period before transmitting control signals (Fullmer & Garcia-Luna-Aceves, 1995).
- **Dual-busy-tone multiple access (DBTMA):** Another significant cause of collision in MACA-based protocols is collision between control packets and data transmission. This problem can be solved by introducing separate channels for control messages, which was proposed in the DBTMA protocol published in 1998 (Haas & Deng, 1998).

ROUTING PROTOCOLS FOR MANET

Routing issues for ad hoc networks with different devices having variable parameters leads to many interesting problems, as evidenced in the literature (Das & Bharghavan, 1997; Dhar, Rieck, Pai, & Kim, 2004; Illyas, 2003; Iwata, Chiang, Pei, Gerla, & Chen, 1999; Liang & Haas, 2000; Perkins, Royer, & Das, 1999; Ramanathan & Streenstrup, 1998; Rieck, Pai, & Dhar, 2002; Toh, 2002; Wu & Li, 2001). This is also validated by industry as well as government efforts such as DoD-sponsored MANET work (Freebersyser & Leiner, 2002). A good network routing protocol may be one that yields the best throughput and response time. However, the very nature of ad hoc networks adds to the requirement for a good routing protocol a set of more, often conflicting, requirements. Accordingly, a good ad hoc routing protocol should also be scalable and reliable. Various routing algorithms and protocols have been introduced in recent years.

Wireless devices are often powered by batteries that have a finite amount of energy. In some ad hoc networks such as sensor networks deployed in a hostile zone, it may not be possible to change a battery once it runs out of energy. As a consequence, the conservation of energy is of foremost concern for those networks. A good ad hoc routing protocol should therefore be energy efficient.

Routing protocols can broadly be classified into four major categories: proactive routing, flooding, reactive routing, and dynamic cluster-based routing (McDonald & Znati, 1999). Proactive routing protocols propagate routing information throughout the network at regular time intervals. This routing information is used to determine paths to all possible destinations. This approach generally demands considerable overhead-message traffic as well as routing-information maintenance. In a flooding approach, packets are sent to all destinations (broadcast) with the expectation that they will arrive at their destination at some point in time. While this means there is no need to worry about routing data, it is clear that for large networks, this generates very heavy traffic, resulting in unacceptably poor overall network performance. Reactive routing maintains path information on a demand basis by utilizing a query-response technique. In this case, the total number of destinations to be maintained for routing information is considerably less than flooding and, hence, the network traffic is also reduced. In dynamic cluster-based routing, the network is partitioned into several clusters, and from each cluster, certain nodes

are elected to be cluster heads. These cluster heads are responsible for maintaining the knowledge of the topology of the network. As it has already been said, clustering may be invoked in a hierarchical fashion.

Some of the specific approaches that have gained prominence in recent years are as follows: The dynamic destination-sequenced distance-vector (DSDV) routing protocol (Johnson & Maltz, 1999), wireless routing protocol (WRP; Murthy & Garcia-Luna-Aceves, 1996), cluster-switch gateway routing (CSGR; Chiang, Wu, & Gerla, 1997), and source-tree adaptive routing (STAR; Garcia-Luna-Aceves & Spohn, 1999) are all examples of proactive routing, while ad hoc on-demand distance-vector routing (AODV; Perkins et al., 1999), dynamic source routing (DSR; Broch, Johnson, & Maltz, 1999), temporally ordered routing algorithm (TORA; Park & Corson, 1997), relative-distance microdiversity routing (RDMAR; Aggelou & Tafazolli, 1999), and signal-stability routing (SSR; Ramanathan & Streenstrup, 1998) are examples of reactive routing. Location-aided routing (LAR; Haas & Liang, 1999) uses location information, possibly via GPS, to improve the performance of ad hoc networks, and global state routing (GSR) is discussed in Chen and Gerla (1998). The power-aware routing (PAR) protocol (Singh, Woo, & Raghavendra, 1998) selects routes that have a longer overall battery life. The zone-Routing protocol (ZRP; Haas & Pearlman, 2000) is a hybrid protocol that has the features of reactive and proactive protocols. Hierarchical state routing (Bannerjee & Khuller, 2001) and cluster-based routing (Amis, Prakash, Vuong, & Huynh, 2000) are examples of dynamic cluster-based routing.

FUTURE TRENDS AND CHALLENGES

MANET will continue to grow in terms of capabilities and applications in consumer as well as commercial markets. There are already quite useful applications of MANET in the military. Currently, it is not just an area of academic research, but also plays an important role in business applications for the future. This trend will continue in the future.

The usefulness of MANET also lies in how this technology will be integrated with the Internet and other wireless technologies like Bluetooth, WLAN, and cellular networks. Another important application of MANET will be in the area of sensor networks, where nodes are not as mobile as MANET but have the essential characteristic of MANET. We will continue to see more and more deployment of sensor networks in various places to collect data and enhance security. So, from that perspective, the future of MANET and its growth looks very promising along with its practical applications.

Although a great deal of work has been done, there are still many important challenges that need to be addressed. We summarize the important issues here.

- **Security and reliability:** Ad hoc networks use wireless links to transmit data. This makes MANET very vulnerable to attack. Although there is some work being done on the security issues of MANET, many important problems and challenges still need to be addressed. With the lack of any centralized architecture or authority, it is always difficult to provide security because key management becomes a difficult problem (Perkins 2002). It is also not easy to detect a malicious node in a multihop ad hoc network and to implement denial of service efficiently. Reliable data communications to a group of mobile nodes that continuously change their locations is extremely important, particularly in emergency situations. In addition, in a multicasting scenario, traffic may pass through unprotected routers that can easily get unauthorized access to sensitive information (as in the case with military applications). There are some solutions that are currently available based on encryption, digital signatures, and so forth in order to achieve authentication and make the MANETs secure, but a great deal of effort is required to achieve a satisfactory level of security. The secure routing protocol (Papadimitratos & Haas, 2002) tries to make MANET more reliable by combating attacks that disrupt the route-discovery process. This protocol will guarantee that the topological information is correct and up to date.
- **Scalability:** Scalability becomes a difficult problem because of the random movement of the nodes along with the limited transmission radius and energy constraints of each node.

- **Quality of service (QoS):** Certain applications require QoS, without which communication will be meaningless. Incorporating QoS in MANET is a nontrivial problem because of the limited bandwidth and energy constraints. The success and future application of MANET will depend on how QoS will be guaranteed in the future.
- **Power management:** Portable handheld devices have limited battery power and often act as nodes in a MANET. They deliver and route packets. Whenever the battery power of a node is depleted, the MANET may cease to operate or may not function efficiently. An important problem is to maximize the lifetime of the network and efficiently route packets.
- **Interoperability:** Integrating MANETs with heterogeneous networks (fixed wireless or wired networks, Internet, etc.) seamlessly is a very important issue. Hosts should be able to migrate from one network to another seamlessly and make pervasive computing a reality.
- **Group membership:** In a MANET, sometimes a new node can join the network, and sometimes some existing nodes may leave the network. This poses a significant challenge for efficient routing management.
- **Mobility:** In MANETs, all the nodes are mobile. Multicasting becomes a difficult problem because the mobility of the nodes creates inefficient multicast trees and an inaccurate configuration of the network topology. In addition, modeling mobility patterns is also an interesting issue. Several researchers have been quite actively investigating this area of research.

CONCLUSION

The growing importance of ad hocs wireless network can hardly be exaggerated as portable wireless devices are now ubiquitous and continue to grow in popularity and in capabilities. In such networks, all of the nodes are mobile, so the infrastructure for message routing must be self-organizing and adaptive. In these networks, routing is an important issue because there is no base station that can be used for broadcasting.

Current and future research will not only address the issues described earlier, but will also try to find new applications of MANET. So far, the research community has been unable to find the killer app using MANET other than in military applications. So, the success of this technology will largely depend on how it will be integrated with the Internet, PANs, and WLANs. MANET will also play an important role in ubiquitous computing, when it will be able to seamlessly integrate with heterogeneous networks and devices, provide various services on demand, and offer secure and reliable communications.

REFERENCES

Aggelou, G., & Tafazolli, R. (1999). RDMAR: A bandwidth-efficient routing protocol for mobile ad hoc networks. *Proceedings of the Second ACM International Workshop on Wireless Mobile Multimedia (WoWMoM)*, Seattle, WA.

Amis, A. D., Prakash, R., Vuong, T. H. P., & Huynh, D. T. (2000). Max-min D-cluster formation in wireless ad hoc networks. *Proceedings of IEEE INFOCOM*, Tel Aviv, Israel.

Bannerjee, S., & Khuller, S. (2001). A clustering scheme for hierarchical control in multi-hop wireless networks. *IEEE Infocom*, Anchorage, AK.

Bharghavan, V., Demers, A., Shenker, S., & Zhang, L. (1994). MACAW: A medium access protocol for wireless LANs. *Proceedings of ACM SIGCOMM '94*, Portland, Oregon.

Broch, J., Johnson, D. & Maltz, D. (1999). The dynamic source routing protocol for mobile ad hoc networks. *IETF, MANET Working Group.* Internet draft '03.

Chen, T.-W. & Gerla, M. (1998). Global state routing: A new routing scheme for ad-hoc wireless networks. *Proceedings IEEE ICC*, Atlanta, Georgia, 171-175.

Chiang, C. C., Wu, H. K., & Gerla, M. (1997). Routing in clustered multihop mobile wireless networks with fading channel. *Proceedings of IEEE Singapore International Conference on Networks*, Singapore.

Das, B., & Bharghavan, V. (1997). Routing in ad-hoc networks using minimum connected dominating

sets. *Proceedings of the IEEE International Conference on Communications (ICC'97)*, 376-380.

Dhar, S., Rieck, M. Q., Pai, S., & Kim, E. J. (2004). Distributed routing schemes for ad hoc networks using d-SPR sets. *Journal of Microprocessors and Microsystems, Special Issues on Resource Management in Wireless and Ad Hoc Mobile Networks, 28*(8), 427-437.

Freebersyser, J., & Leiner, B. (2002). A DoD perspective on mobile ad hoc networks. In C. Perkins (Ed.), *Ad hoc networking.* Upper Saddle River, NJ: Addison Wesley.

Fullmer, C., & Garcia-Luna-Aceves, J. J. (1995). Floor acquisition multiple access (FAMA) for packet radio networks. *Computer Communication Review, 25*(4), 262-273.

Garcia-Luna-Aceves, J. J., & Spohn, M. (1999). Source tree adaptive routing in wireless networks. *Proceedings of IEEE ICNP,* Toronto, Canada.

Haas, Z., & Deng, J. (1998). Dual busy tone multiple access (DBTMA): A new medium access control for packet radio networks. *IEEE 1998 International Conference on Universal Personal Communications*, Florence, Italy.

Haas, Z.J. & Liang, B. (1999). Ad hoc location management using quorum systems. *ACM/IEEE Transactions on Networking, 7*(2), 228-240.

Haas, Z.J. & Pearlman, M. (2000). The zone routing protocol (zpc) for ad hoc networks. IETF, MANET Working Group, Internet draft '03. Retrieved from *http://www.ics.uci.edu/~atm/adhoc/paper-collection/haas-draft-ietf-manet-zone-zrp-00.txt*

Illyas, M. (2003). *The handbook of ad hoc wireless networks.* Boca Raton, FL: CRC Press.

Iwata, A., Chiang, C.-C., Pei, G., Gerla, M., & Chen, T. W. (1999). Scalable routing strategies for ad hoc wireless networks. *IEEE Journal on Selected Areas in Communications*, *7*(8), 1369-1379.

Johnson, D. B, & Maltz, D. A. (1999). *The dynamic source routing protocol for mobile ad hoc networks* (IETF draft). Retrieved from *http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-03.txt*

Karn, P. (1992). MACA: A new channel access method for packet radio. *Proceedings of the Ninth ARRL/CRRL Amateur Radio Computer Networking Conference*, 134-140.

Liang, B., & Haas, Z. J. (2000). Virtual backbone generation and maintenance in ad hoc network mobility management. *Proceedings of IEEE Infocom, 5*, 1293-1302.

McDonald, A. B., & Znati, T. (1999). A mobility-based framework for adaptive clustering in wireless ad-hoc networks. *IEEE Journal on Selected Areas in Communications, 17*(8), 1466-1487.

Murthy, S., & Garcia-Luna-Aceves, J. J. (1996). An efficient routing protocol for wireless networks. *ACM Mobile Networks and Applications, 1*(2), 183-197.

Papadimitritratos, P., & Haas, Z. (2002). Secure routing for mobile ad hoc networks. *Proceedings of CNDS,* San Antonio, Texas.

Park, V.D. & Corson, M.S. (1997). A highly adaptive distributed routing algorithm for mobile wireless networks. *Proceedsings IEEE INFOCOM*, 1405-1413.

Perkins, C. (2002). *Ad hoc networking.* Upper Saddle River, NJ: Prentice Hall.

Perkins, C. E., Royer, E. M., & Das, S. R. (1999). *Ad hoc on-demand distance vector routing* (IETF draft). Retrieved from *http://www.ietf.org/internet-drafts/draft-ietf-manet-aodv-04.txt*

Ramanathan, R., & Streenstrup, M. (1998). Hierarchically organized, multi-hop mobile wireless networks for quality-of-service support. *Mobile Networks and Applications, 3*, 101-119.

Rieck, M. Q., Pai, S., & Dhar, S. (2002). Distributed routing algorithms for wireless ad hoc networks using d-hop connected d-hop dominating sets. *Proceedings of the Sixth International Conference on High Performance Computing: Asia Pacific*, 443-450.

Singh, S., Woo, M., & Raghavendra, C. S. (1998). Power-aware routing in mobile ad hoc networks. *Proceedings of ACM/IEEE Mobicom*, 181-190.

Toh, C.-K. (2002). *Ad hoc wireless mobile networks*. Upper Saddle River, NJ: Prentice Hall Inc.

Weiser, M. (1993). Some computer sciences issues in ubiquitous computing. *Communications of the ACM, 36*(7), 75-84.

Wu, J., & Li, H. (2001). A dominating-set-based routing scheme in ad hoc wireless networks. *Telecommunication Systems, 18*(1-3), 13-36.

Wu, J., & Stojmenovic, I. (2004, February). Ad hoc networks. *IEEE Computer*, 29-31.

KEY TERMS

CSMA: Carrier-sense multiple access is a media-access control (MAC) protocol in which a node verifies the absence of other traffic before transmitting on a shared physical medium, such as an electrical bus or a band of electromagnetic spectrum. Carrier sense describes the fact that a transmitter listens for a carrier wave before trying to send. That is, it tries to detect the presence of an encoded signal from another station before attempting to transmit. Multiple access describes the fact that multiple nodes may concurrently send and receive on the medium.

GPS: It stands for Global Positioning System. It is an MEO (medium earth orbit) public satellite navigation system consisting of 24 satellites used for determining one's precise location and providing a highly accurate time reference almost anywhere on Earth.

MAC: Media-access control is the lower sublayer of the OSI (open systems interconnection reference model) data-link layer: the interface between a node's logical link control and the network's physical layer. The MAC sublayer is primarily concerned with breaking data up into data frames, transmitting the frames sequentially, processing the acknowledgment frames sent back by the receiver, handling address recognition, and controlling access to the medium.

MANET: A mobile ad hoc network is a system of wireless mobile nodes that dynamically self-organize in arbitrary and temporary topologies.

Peer-to-Peer Network: A peer-to-peer (or P2P) computer network is any network that does not have fixed clients and servers, but a number of *peer* nodes that function as both clients and servers to the other nodes on the network. This model of network arrangement is contrasted with the client-server model. Any node is able to initiate or complete any supported transaction. Peer nodes may differ in local configuration, processing speed, network bandwidth, and storage quantity.

Routing Protocol: Routing protocols facilitate the exchange of routing information between networks, allowing routers to build routing tables dynamically.

Ubiquitous Computing: This is a term describing the concept of integrating computation into the environment rather than having computers that are distinct objects. Promoters of this idea hope that embedding computation into the environment will enable people to move around and interact with computers more naturally than they currently do.

Mobile Agents

Kamel Karoui
Institut National des Sciences Appliquées de Tunis, Tunisia

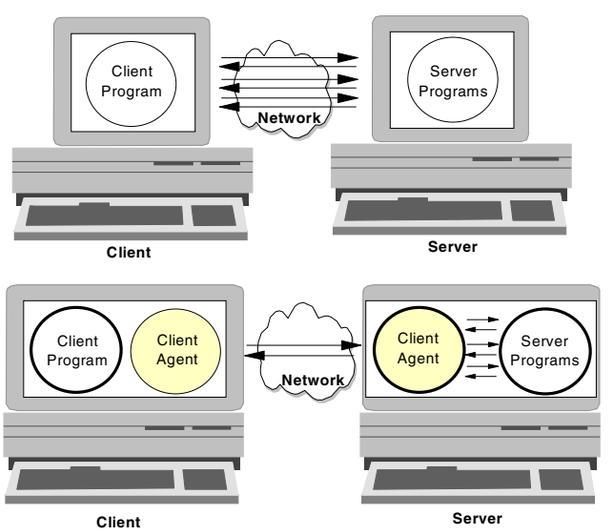
INTRODUCTION

The concept of mobile agent is not new; it comes from the idea of *OS process migration* firstly presented by *Xerox* in the 1980's. The term *mobile agent* was introduced by White & Miller (1994), which supported the mobility as a new feature in their programming language called *Telescript*.

This new research topic has emerged from a successful meeting of several sub-sciences: computer networks, software engineering, object-oriented programming, artificial intelligence, human-computer interaction, distributed and concurrent systems, mobile systems, telematics, computer-supported cooperative work, control systems, mining, decision support, information retrieval and management, and electronic commerce. It is also the fruit of exceptional advances in distributed systems field (Hirano 1997; Holder, Ben-Shaul, & Gazit 1999; Lange et al., 1999).

The main idea of the mobile agent technology is to replace the old approach of the client-server Remote Procedure Call (RPC) paradigm, by a new one consisting of transporting and executing programs around a network. The results of the programs execution are then returned back to the sending entity. Figure 1 illustrates this new approach.

Figure 1. RPC vs. mobile agent approach



Mobile agents are dynamic, non-deterministic, unpredictable, proactive, and autonomous entities. They can decide to exercise some degree of activities without being invoked by external entities. They can watch out for their own set of internal responsibilities. Agents can interact with their environment and other entities. They can support method invocation as well as more complex degree of interaction as for example the observable events reaction within their environment. They can decide to move from one server to another in order to accomplish the system global behavior.

BACKGROUND

As the information technology moves from a focus on the individual computer system to a situation in which the real power of computers is realized through distributed, open and dynamic systems, we are faced with new technological challenges. The characteristics of dynamic and open environments in which heterogeneous systems must interact require improvements on the traditional computing models and paradigms. It is clear that these new systems need some degree of intelligence, autonomy, mobility, and so on. The mobile agent concept is one of the new system environment that has emerged from this need. Several researches have proposed a definition of mobile agents (Bradshaw, Greaves, Holmback, Jansen, Karygiannis, Silverman, Suri, & Wong, 1999; Green & Somers, 1997; White 1997). Until now, there is neither standard nor a unique consensus on a unique definition. In general, a mobile agent can be defined using its basic attributes: the mobility, the intelligence and the interactivity. Based on these attributes, we can propose the following definition:

A mobile agent is a computational entity which acts on behalf of other entities in an intelligent way

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

(autonomy, learning, reasoning, etc.). It performs its tasks in software open and distributed environment with some level of mobility, co-operation, proactivity, and/or reactiveness.

This attributes based definition gives an abstract view of what a mobile agent does, but it doesn't present how it does it. This definition doesn't mean that mobility, interactivity, and intelligence are the unique attributes of mobile agents. Effectively, a large list of other attributes exists such as: application field, communication, delegation, and so on.

This definition shows that a mobile agent doesn't exist without a software environment called a mobile agent environment (see Figure 2).

AGENT CLASSIFICATION

According to the literature (Frankllin & Graesser, 1996), agents, and especially mobile agents, can be classified using the three agent basic attributes depicted in Figure 3.

- The first agent attribute is mobility, so an agent can be static or mobile.
- The second attribute is intelligence; an agent can be characterized by its abilities of reasoning, planning, learning, and so on.
- Interaction is the third agent attribute. Agents can have different kinds of interactions. This category of agents contains the agents that: do not interact at all, interact with users, interact with applications, and interact with other agents.

There are of course many other classification methods (Frankllin & Graesser, 1996). For example, we can classify agents according to the task they perform, for example, information gathering agents or e-mail filtering agents.

MOBILE AGENT ADVANTAGES

Using mobile agents is not the unique way to solve some class of problems, alternative solutions exists. However, for some class of problems and applications, we believe that mobile agent technology is more adapted than classical methods. For example, in managing large scale intranet, where we must continuously, install, update, and customize software for different users without bringing the server down. In the following we present three types of application domains where it is better to use mobile agent technology:

- Data-intensive application where the data is remotely located. Here, agents are sent in order to process and retrieve data.
- Disconnected computing application where agents are launched by an appliance. For example, shipping an agent from a cellular phone to a remote server.
- Application where we need to extend the server behavior by sending agents that can represent permanently or not the server in different location (host or server).

Figure 2. Mobile agent environment

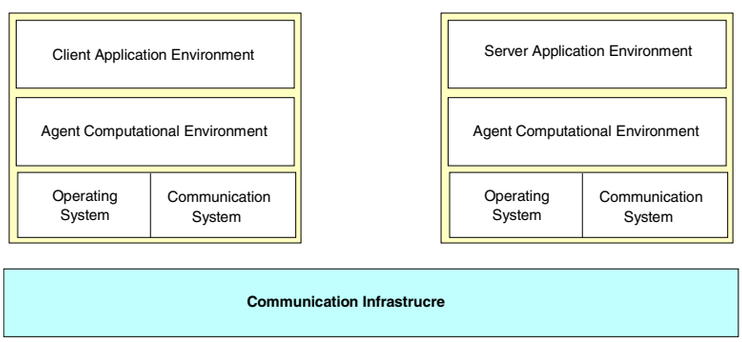
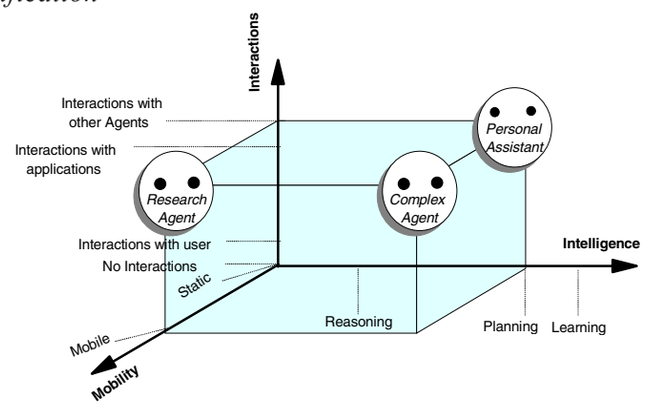


Figure 3. Agent classification



In the following we present a list of the main advantages of mobile agent's technology:

- Efficiency: mobile agents consume fewer network resources.
- Reduction of the network traffic: mobile agents minimize the volume of interactions by moving and executing programs on special host servers.
- Asynchronous autonomous interactions: mobile agents can achieve tasks asynchronously and independently of the sending entity.
- Interaction with real-time entities: for critical systems (nuclear, medical, etc.) agents can be dispatched from a central site to control local real-time entities and process directives from the central controller.
- Dynamic adaptation: mobile agents can dynamically react to changes in its environment.
- Dealing with vast volumes of data: by moving the computational to the sites containing a large amount of data instead of moving data, we can reduce the network traffic.
- Robustness and fault tolerance: by its nature, a mobile agent is able to react to multiple situations, especially faulty ones. This ability makes the systems based on mobile agents fault tolerant.
- Support for heterogeneous environments: mobile agents are generally computer and network independent, this characteristic allows their use in a heterogeneous environment.

MOBILE AGENT DISADVANTAGES

In the following we present a list of the major problems for mobile agent approach:

- Security is one of the main concerns of the mobile agent approach. The issue is how to protect agent from malicious hosts and inversely how to protect hosts from mobile agents. The main researchers' orientation is to isolate the agent execution environment from the host critical environment. This separation may limit the agent capabilities of accessing the desired data and from accomplishing its task.
- Another big problem of the mobile agent approach is the lack of standardization. In the recent years, we have seen the development of many mobile agent systems based on several slightly different semantics for mobility, security, and communication. This will restrict the developers to small applications for particular software environments.
- Mobile agents are not the unique way to solve major class of problems, alternative solutions exists: messaging, simple datagram, sockets, RPC, conversations, and so on. There are neither measurement methods nor criteria that can help developer choose between those methods. Until now there is no killer application that uses the mobile agent approach.
- Mobile agents can achieve tasks asynchronously and independently of the sending entity. This can be an advantage for batch applications and disadvantage for interactive applications.

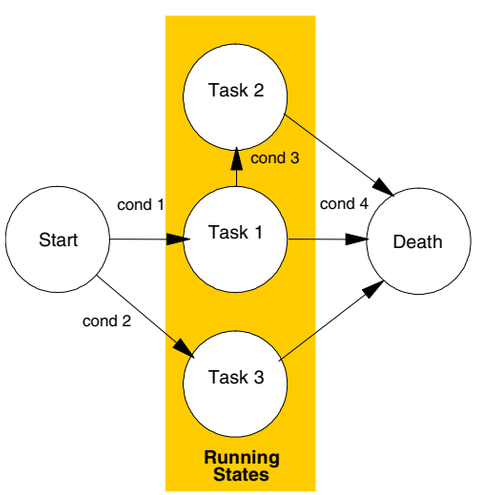
MOBILE AGENT MODELS

A successful mobile agent system should be designed based on the following six models. The implementa-

tion of these models depends on the agent construction tools.

- **Agent model:** It defines the intelligent part (autonomy, reasoning, learning, etc.) of the agent internal structure.
- **Computational model:** It defines how the agent executes its self when it is in its running states (see Figure 4). In general, this model is represented by a finite state machine or an extended finite state machine (Karoui, Dssouli, & Yevtushenko 1997).
- **Security model:** This model describes the different approach of the security part of the system. In general, there are two main security concerns, protection of hosts from malicious agents and protection of agents from malicious hosts.
- **Communication model:** It presents how the agents communicate and interacts with other agents of the system.
- **Navigation model:** This model deals with the mobility in the system. It describes how an agent is transported from one host to another.
- **Life-cycle model:** Each agent can be characterized by a life cycle. The life cycle starts from the agent creation state *Start*, and ends in the death state *Death*. The intermediate states depend on the nature of the mission. Those last states are called running states (see Figure 4).

Figure 4. Agent life cycle model



AGENT CONSTRUCTION TOOLS

Several mobile agent construction tools have appeared since 1994. Most of them are built on top of the Java system or the Tcl/tk system (Morisson & Lehenbauer 1992). Table 1 provides a survey of some currently available agent construction tools. Although each of these tools supports different levels of functionality, they each attempt to address the same problem: namely, enabling portions of code to execute on different machines within a wide-area network. Many research groups are now focusing on Java as the development language of choice thanks to its portability and code mobility features.

One feature that all of these mobile agent construction tools have currently failed to address is in defining a domain of applicability; they all concentrate on the mobility of agents rather than the integration of agents with information resources.

MOBILE AGENT-BASED SYSTEM EXAMPLE

As an example of multi-agent and mobile agent systems, we present an application in telemedicine that we have developed in previous works (Karoui, Loukil, & Sounbati 2001; Karoui & Samouda 2001). The idea from proposing such system starts from the statistics about health care system of a small country. We have seen that this system suffers from two main weaknesses: insufficiency of specialists and bad distribution of the specialists over the country. Thus, we thought about a system which is able to provide, to a non-expert practitioner (Physician), the appropriate computerized or not help of a distant expert. The system is influenced by the following set of constraints and considerations:

1) Before asking for a help of a distant expert, the system should be able to proceed a multilevel automatic diagnose in order to refine, classify, and document the case.
2) The non-expert site of our system should be able to learn from previous experiences and the diagnosed cases by specialists.
3) The responses should not exceed a limit of time specified by the requestor on the basis of the case emergency.

Table 1. Mobile agent construction tools

Product	Company	Lang.	Description
AgenTalk	NTT/Ishida	LISP	Multiagent Coord.
Agentx	International Knowledge Systems	Java	Agent Development Environment
Aglets	IBM Japan	Java	Mobile Agents
Concordia	Mitsubishi Electric	Java	Mobile Agents
DirectIA SDK	MASA - Adaptive Objects	C++	Adaptive Agents
Gossip	Tryllian	Java	Mobile Agents
Grasshopper	IKV++	Java	Mobile Agents
iGEN™	CHI Systems	C/C++	Cognitive Agent
JACK Intelli Agents	Agent Oriented Software Pty. Ltd.	JACK	Agent Development Environment
JAM	Intelligent Reasoning Systems	Java	Agent Architecture
LiveAgent	Alcatel	Java	Internet Agent
AgentTcl	Dartmouth College	Tcl/tk	Mobile Agents
MS Agent	Microsoft Corp.	Active X	Interface creatures

4) The expert can refuse to respond to a query.
5) In order to facilitate and accelerate the expert diagnosis, the information related to a query and sent to the experts should be as complete as possible.
6) For security purposes, the system should ensure the authentication of both the requestor and the advisor, and also the integrity and confidentiality of the interchanged data.
7) The system should be easy to extend and to maintain.

Taking into account these requirements, we present here after how the system works. First of all, our system is composed of a set of medical sites; each of them has a server connected to a telemedicine network. This later can be either a private network or the Internet. In each medical site we have at least one physician able to collect patient symptoms. When a patient goes for a consultation, we cannot insure that in his local medical center there has the appropriate expert for his disease. In case of expert deficiency, the local physician collects the symptoms through a guided computerized user interface, and a multilevel diagnoses process. In the following, we explain the four-level diagnosis process which is composed of two human diagnoses (levels 1 and 4) and two computerized automatic diagnosis (levels 2 and 3).

1. **The first level diagnosis:** The physician who collects the symptoms can propose a diagnosis. This diagnosis will be verified by a computerized process called the second level diagnosis.
2. **The second level diagnosis:** The local system analyzes automatically the collected symptoms. If the system detect a disease, it automatically informs the physician giving him all the information used to reach such diagnosis (used rules and symptoms). This diagnostic may be different from the one given by the physician himself (*first level diagnosis)*. The system then asks the physician if he wants to confirm this diagnosis by getting the advice of an expert. If yes, a request is sent to a set of experts chosen automatically by the system. The request is represented by mobile agents sent to distant servers. The request contains all the information needed to get the right diagnosis.
3. **The third level diagnosis:** When the distant servers receive the request, each one of them verifies automatically the correctness of the information used in order to produce and send back to the requester (through the mobile agent) a computerized *third level diagnosis*. If this information is not correct (not complete or bad rules), the request is returned back to the sender asking the local system for more special information or symptoms about the case.

4. **The fourth level diagnosis:** If the information contained in the request is correct, but the expert server site cannot produce a computerized diagnostic (*third level diagnosis*). It is presented to a human expert who will analyze, give his diagnosis about the case and take the necessary actions.

For the system performance, we suppose that the non expert part learns (self learning) from its previous experience. So, for a given case, the system starts with a minimal amount of information about diseases, then from the multilevel diagnosis process (specially the third level diagnosis) the system will automatically update its diagnostic rules and databases.

CONCLUSION

Agent-oriented approach is becoming popular in the software development community. In the future, agent technology may become be a dominant approach. The agent-based way of thinking brings a useful and important perspective for system development.

Recent years have seen the development of many mobile agent systems based on several slightly different semantics for mobility, security, communication, and so on. We need now to start the process of choosing the best ideas from the huge number of the proposed approaches and identify the situations where those approaches are useful and may be applied. In order to achieve this goal, we need some quantitative measurements of each kind of mobility communication and security methods. This will automatically result in a kind of standardization.

REFERENCES

Bradshaw, J.M., Greaves, M., Holmback, H., Jansen, W., Karygiannis, T., Silverman, B., Suri, N., & Wong, A. (1999). Agents for the masses: Is it possible to make development of sophisticated agents simple enough to be practical? *IEEE Intelligent Systems*. 53-63.

Franklin, S. & Graesser, A. (1996). Is it an agent, or just a program?: A Taxonomy for Autonomous Agents. *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages.* Springer-Verlag.

Fuggetta, A., Picco, G.P., & Vigna, G. (1998). Understanding Code Mobility. *IEEE Transactions on Software Engineering*, *24*(5).

Green, S. & Somers, F. (1997). Software Agents: A review. Retrieved August 5 1998, from *http://www.cs.tcd.ie/research_groups/aig/iag/pubreview/*

Hirano, S. (1997). HORB: Distributed Execution of Java Programs, *Worldwide Computing and Its Applications'97, Springer Lecture Notes in Computer Science, 1274*, 29-42.

Holder, O., Ben-Shaul, I., & Gazit, H. (1999). System Support for Dynamic Layout of Distributed Application. *Proceedings of the 21 st International Conference on Software Engineering (ICSE'99)*, 163- 173.

IBM (1998). Aglets software development kit. Retrieved June 4, 1999. From *http://www.trl.ibm.co.jp/aglets/*

Karoui, K., Dssouli, R., & Yevtushenko, N. (1997). Design For testability of communication protocols based on SDL. *Eighth SDL FORUM 97, Evry France*.

Karoui, K., Loukil, A., & Sonbaty, Y. (2001). Mobile agent hybrid route determination framework for healthcare telemedicine systems, *ISC 2001,* Tampa Bay USA.

Karoui, K., Samouda, R., & Samouda, M. (2001). Framework for a telemedicine multilevel diagnose system. *IEEE EMBS'2001, Vol. 4,* Istanbul, Turkey, 3508-3512.

Kotz, D., Gray, R., Nog, S., Rus, D., Chawla, S., & Cybenko, G. (1997). Agent TCL: Targeting the needs of mobile computers. *IEEE Internet Computing*, *1*(4).

Lange, D. et al. (1999). Seven good reasons for mobile agents. *Communications of the ACM*, *42*(3), 88-89.

Lange, D. & Oshima, M. (1998). *Programming and deploying java mobile agents with aglets.* Addison - Wesley.

Morisson, B. & Lehenbauer, K. (1992). Tcl and Tk: Tools for the system administration.administration. *Proceedings of the Sixth System Administration Conference,*225-234.

White, J. (1997). *Mobile agent*. J.M. Bradshaw (Ed.), Software Agents. Cambridge, MA: The AAAI Press/The MIT Press, 437-472.

White, J.E. (1994). *Telescript technology: The foundation for the electronic marketplace*. Mountain View, CA: General Magic, Inc.

KEY TERMS

Agent: A computational entity which acts on behalf of other entities.

Agent Attributes: An agent can be classified using some of its characteristics called attributes. An agent has three basic attributes: mobility, intelligence, and interaction.

Client-Server Model: A client-server model defines a basis for communication between two programs called respectively the client and the server. The requesting program is a client and the service-providing program is the server.

Intelligent Agent: An agent who acts in an intelligent way (autonomy, learning, reasoning, etc.).

Mobile Agent: An intelligent agent who performs its tasks with some level of mobility, cooperation, proactivity, and/or reactiveness.

Multiagent System: A system composed of agents interacting together in order to achieve the system common task or behaviour.

RPC: Remote Procedure Call is one way of communication in a client server model. The client and the server are located in different computers in a network. An RPC is a synchronous operation requiring the requesting (client) to pass by value all the needed parameters to the server then the client is suspended until the server returns back the associated results.

Mobile Commerce Security and Payment

Chung-wei Lee
Auburn University, USA

Weidong Kou
Xidian University, PR China

Wen-Chen Hu
University of North Dakota, USA

INTRODUCTION

With the introduction of the World Wide Web (WWW), electronic commerce has revolutionized traditional commerce and boosted sales and exchanges of merchandise and information. Recently, the emergence of wireless and mobile networks has made possible the extension of electronic commerce to a new application and research area: mobile commerce, which is defined as the exchange or buying and selling of commodities, services or information on the Internet through the use of mobile handheld devices. In just a few years, mobile commerce has emerged from nowhere to become the hottest new trend in business transactions. Mobile commerce is an effective and convenient way of delivering electronic commerce to consumers from anywhere and at any time. Realizing the advantages to be gained from mobile commerce, companies have begun to offer mobile commerce options for their customers in addition to the electronic commerce they already provide (The Yankee Group, 2002).

Regardless of the bright future of mobile commerce, its prosperity and popularity will be brought to a higher level only if information can be securely and safely exchanged among end systems (mobile users and content providers). Applying the security and payment technologies for electronic commerce to mobile commerce has been proven a futile effort because electronic commerce and mobile commerce are based on different infrastructures (wired vs. wireless). A wide variety of security procedures and payment methods, therefore, have been developed and applied to mobile commerce. These technologies are extremely diverse and complicated. This article provides a comprehensive overview of mobile commerce security and payment methods.

BACKGROUND

Mobile security is a crucial issue for mobile commerce. Without secure commercial information exchange and safe electronic financial transactions over mobile networks, neither service providers nor potential customers will trust mobile commerce systems. From a technical point of view, mobile commerce over wireless networks is inherently insecure compared to electronic commerce over the Internet (Pahlavan & Krishnamurthy, 2002). The reasons are as follows:

- Reliability and integrity: Interference and fading make the wireless channel error-prone. Frequent handoffs and disconnections also degrade the security services.
- Confidentiality/privacy: The broadcast nature of the radio channel makes it easier to tap. Thus, communication can be intercepted and interpreted without difficulty if no security mechanisms such as cryptographic encryption are employed.
- Identification and authentication: The mobility of wireless devices introduces an additional difficulty in identifying and authenticating mobile terminals.
- Capability: Wireless devices usually have limited computation capability, memory size, com-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

munication bandwidth and battery power. This will make it difficult to utilize high-level security schemes such as 256-bit encryption.

Mobile commerce security is tightly coupled with network security. The security issues span the whole mobile commerce system, from one end to the other, from the top to the bottom network protocol stack, from machines to humans. Therefore, many security mechanisms and systems used in the Internet may be involved. In this article we focus only on issues exclusively related to mobile/wireless technologies. On a secure mobile commerce platform, mobile payment methods enable the transfer of financial value and corresponding services or items between different participators without factual contract. According to the amount of transaction value, mobile payment can be divided into two categories. One is micro-payment, which defines a mobile payment of approximately $10 or less (ComputerWorld, 2000), often for mobile content such as video downloads or gaming. The other is macro-payment, which refers to larger-value payments.

MOBILE COMMERCE SECURITY

Mobile commerce transactions can be conducted on the infrastructure of wireless cellular networks as well as wireless local area networks. Lacking a unified wireless security standard, different wireless networking technologies support different aspects and levels of security features. We thus discuss some popular wireless network standards and their corresponding security issues.

Wireless Cellular Network and Security

In addition to voice communication, cellular network users can conduct mobile commerce transactions through their well-equipped cellular phones. Currently, most of the cellular wireless networks in the world follow second-generation (2G, 2.5G) standards. Examples are the global system for mobile communications (GSM) and its enhancement, general packet radio service (GPRS). GPRS can support data rates of only about 100 kbps, and its upgraded version – enhanced data for global evolution (EDGE) – is capable of supporting 384 kbps. It is expected that third-generation (3G) systems will dominate wireless cellular services in the near future. The two main standards for 3G are Wideband CDMA (WCDMA), proposed by Ericsson, and CDMA2000, proposed by Qualcomm. The WCDMA system can inter-network with GSM networks and has been strongly supported by the European Union, which calls it the Universal Mobile Telecommunications System (UMTS). CDMA2000 is backward-compatible with IS-95, which is widely deployed in the United States.

GSM Security

The Subscriber Identity Module (SIM) in the GSM contains the subscriber's authentication information, such as cryptographic keys, and a unique identifier called international mobile subscriber identity (IMSI). The SIM is usually implemented as a smart card consisting of microprocessors and memory chips. The same authentication key and IMSI are stored on GSM's network side in the authentication center (AuC) and home location register (HLR), respectively. In GSM, short messages are stored in the SIM and calls are directed to the SIM rather than the mobile terminal. This feature allows GSM subscribers to share a terminal with different SIM cards. The security features provided between the GSM network and mobile station include IMSI confidentiality and authentication, user data confidentiality and signaling information element confidentiality. One of the security weaknesses identified in GSM is the one-way authentication. That is, only the mobile station is authenticated; the network is not. This can pose a security threat, as a compromised base station can launch a "man-in-the-middle" attack without being detected by mobile stations.

UMTS Security

UMTS is designed to reuse and evolve from existing core network components of the GSM/GPRS and fix known GSM security weaknesses such as the one-way authentication scheme and optional encryption. Authentication in UMTS is mutual and encryption is mandatory (unless specified otherwise) to prevent message replay and modification. In addition, UMTS employs longer cryptographic keys and newer cipher algorithms that make it more secure than GSM/GPRS.

Wireless Local Area Network and Security

Among popular wireless local area networks (WLANs), Bluetooth technology supports very limited coverage range and throughput. Thus, it is only suitable for applications in personal area networks (PANs). In many parts of the world, the IEEE 802.11b (Wi-Fi) system is the dominant WLAN and is widely deployed in offices, homes and public spaces such as airports, shopping malls and restaurants. Even so, many experts predict that with much higher transmission speeds, 802.11g will replace 802.11b in the near future.

Wi-Fi Security

The security of the IEEE 802.11 WLAN standard is provided by a data-link-level protocol called Wired Equivalent Privacy (WEP). When it is enabled, each mobile host has a secret key shared with the base station. The encryption algorithm used in WEP is a synchronous stream cipher based on RC4. The ciphertext is generated by XORing the plain text with an RC4-generated keystream. However, recently published literature has discovered weaknesses in RC4 (Borisov, Goldberg & Wagner, 2001; Fluhrer, Martin & Shamir, 2001; Stubblefield, Ioannidis & Rubin, 2002). The new version, 802.11i (Cam-Winget, Moore, Stanley & Walker, 2002), is expected to have better security by employing an authentication server that separates authentication process from the AP.

Bluetooth Security

Bluetooth provides security by using frequency hopping in the physical layer, sharing secret keys (called passkeys) between the slave and the master, encrypting communication channels and controlling integrity. Encryption in Bluetooth is a stream cipher called "E_0," while for integrity control a block cipher called "SAFER+" is used. However, "E_0" has potential weaknesses, as described in Jakobsson and Wetzel (2001) and Biryukov, Shamir and Wagner (2000), and "SAFER+" is slower than the other similar symmetric-key block ciphers (Tanenbaum, 2002). Security in Bluetooth networks can be strengthened by employing service-level functions such as the Security Manager (Ma & Cao, 2003).

WAP and Security

Beyond the link-layer communication mechanisms provided by WLANs and cellular networks, the Wireless Application Protocol (WAP) is designed to work with all wireless networks. The most important component in WAP is probably the Gateway, which translates requests from the WAP protocol stack to the WWW stack, so they can be submitted to Web servers. For example, requests from mobile stations are sent as a URL through the network to the WAP Gateway; responses are sent from the Web server to the WAP Gateway in HTML and are then translated to Wireless Markup Language (WML) and sent to the mobile stations.

WAP security is provided through the Wireless Transport Layer Security (WTLS) protocol (in WAP 1.0) and IETF standard Transport Layer Security (TLS) protocol (in WAP 2.0). They provide data integrity, privacy and authentication. One security problem, known as the "WAP Gap," is caused by the inclusion of the WAP gateway in a security session. That is, encrypted messages sent by end systems might temporarily become clear text on the WAP gateway when messages are processed. One solution is to make the WAP gateway resident within the enterprise (server) network (Ashley, Hinton & Vandenwauver, 2001), where heavyweight security mechanisms can be enforced.

MOBILE COMMERCE PAYMENT

There are four players in a mobile payment transaction. The mobile consumer (MC) subscribes to a product or service and pays for it via mobile device. The content provider/merchant (CP/M) provides the appropriate digital content, physical product or service product to the consumer. The payment service provider (PSP), which may be a network operator, financial institution or independent payment vendor, controls the payment process. The trusted third party (TTP) administers the authentication of transaction parties and the authorization of the payment settlement. In fact, the different roles can be merged into one organization; for example, a network bank, which is capable of acting as CP/M, PSP and TTP at the same time. In a more general sense, a PSP and TTP can be performed by the same organization.

Mobile Payment Scenarios

Content Download

In this scenario, consumers order the content they want to download from a content provider. The content provider then initiates the charging session, asking the PSP for authorization. The PSP authorizes the CP/M, and then the download starts. The transaction can be settled by either a metered or pricing model. The metered content includes streaming services. The consumers are charged according to the metered quantity of the provided service; for example, interval, data volume or gaming sessions. In a pricing model, consumers re charged according to the items downloaded completely. A content purchase is also available via PC Internet connection, where the mobile device will be used to authorize the payment transaction and authenticate the content user.

Point of Sale

In this scenario, services or the sale of goods are offered to the mobile user on the point-of-sale location instead of a virtual site; for example, a taxi service. The merchant (e.g., the taxi driver) will initiate payment at the point of sale. The PSP asks the mobile user to directly authorize the transaction via SMS pin or indirectly via the taxi driver through a wireless Bluetooth link. The process is also applicable to a vending-machine scenario.

Content on Device

In this payment scenario, users have the content preinstalled in their mobile device, but should be granted a license to initiate the usage of the content; for example, the activation of an on-demand gaming service. The license varies with usage, duration or number of users, and determines the value that the consumer should pay for the desired content.

Mobile Payment Methods

Out-of-Band Payment Method

In the "out-of-band" model, content and operation signals are transmitted in separate channels; for example, credit card holders may use their mobile devices to authenticate and pay for a service they consume on the fixed-line Internet or interactive TV. This model usually involves a system controlled by a financial institution, sometimes collaborating with a mobile operator. There are two typical cases:

Financial Institutions

A great number of banks are conducting research to turn the individual mobile into a disbursing terminal. Payments involved in the financial transaction are usually macro-payments. Various methods can be deployed to ensure the authentication of payment transaction. In credit card payments, a dual-slot phone is usually adopted. Other approaches include PIN authentication via SIM toolkit application and also the use of a digital signature based on a public key infrastructure (PKI) mechanism that demands the 2.5G (or higher) technology.

Reverse-Charge/Billed SMS

In reverse-billed premium-rate SMS, the CP/M deliver content to mobile telephone handsets (ICSTIS, n.d.). Customers subscribe to a service and are charged for the messages they receive. This payment model allows consumers to use SMS text messages to pay for access to digital entertainment and content without being identified. In this application, however, it is the SMS message receiver who is charged, instead of the sender of the SMS message. There are a considerable number of vendors who offer the reverse-charge/billed MSM service payment models.

In-Band Payment Method

In this method, a single channel is deployed for the transfer of both content and operation signals. A chargeable WAP service over GPRS is of this kind. Two models of this in-bank payment are in use; namely, subscription models and per-usage payment models, with the amount of the payment usually being small; that is, micro payments. In-band transactions include applications such as video streaming of sports highlights or video messaging.

Proximity

Proximity payments involve the use of wireless technologies to pay for goods and services over short distances. Proximity transactions develops the potential of mobile commerce; for example, using a mobile device to pay at a point of sale, vending machine, ticket machine, market, parking and so forth. Through short-range messaging protocols such as Bluetooth, infrared, RFID or contactless chip, the mobile device can be transformed to a sophisticated terminal able to process both micro and macro payments (DeClercq, 2002).

Mobile Payment Standardization

Current mobile payment standardization has mainly been developed by several organizations, as follows:

- Mobey Forum (2002): Founded by a number of financial institutions and mobile terminal manufacturers, Mobey Forum's mission is to encourage the use of mobile technology in financial services.
- Mobile Payment Forum (2002): This group is dedicated to developing a framework for standardized, secure and authenticated mobile commerce using payment card accounts.
- Mobile electronic Transactions (MeT) Ltd. (2002): This group's objective is to ensure the interoperability of mobile transaction solutions. Its work is based on existing specifications and standards, including WAP.

FUTURE TRENDS

Mobile telecommunications has been so successful that the number of mobile subscribers had risen to 1 billion worldwide by the end of 2002. It is estimated that 50 million wireless phone users in the United States will use their handheld devices to authorize payment for premium content and physical goods at some point during 2006. This represents 17% of the projected total population and 26% of all wireless users (Reuters, 2001). Accompanying the increase in subscriptions, there are evolutions in more sophisticated devices, encouraging the emergence of new applications that include enhanced messaging services (EMS) and multimedia messaging services (MMS). In these applications, consumers have more options, such as the download of images, streaming video and data files, as well as the addition of global positioning systems (GPS) in mobile phones, which will facilitate location-based and context-aware mobile commerce, and furthermore provide more feasibility to mobile payment methods.

For security issues in the wireless networking infrastructure, testing and developing new secure protocols at all network layers are very important to mobile commerce's prosperity. For example, newly standardized IEEE 802.11i requires stringent evaluation in the real world. At the transport layer, TCP can be modified to avoid WAP security flaws (Juul & Jorgensen, 2002). In addition, low-complexity security protocols and cryptographic algorithms are needed to cope with the constrained computation power and battery life in a typical wireless handheld device.

CONCLUSION

It is widely acknowledged that mobile commerce is a field of enormous potential. However, it is also commonly admitted that the development in this field is constrained. There are still considerable barriers waiting to be overcome. Among these, mobile security and payment methods are probably the biggest obstacles. Without secure commercial information exchange and safe electronic financial transactions over mobile networks, neither service providers nor potential customers will trust mobile commerce.

Mobile commerce security is tightly coupled with network security; however, lacking a unified wireless security standard, different wireless technologies support different aspects and levels of security features. This article, therefore, discussed the security issues related to the following three network paradigms: 1) wireless cellular networks; 2) wireless local area networks; and 3) WAP.

Among the many themes of mobile commerce security, mobile payment methods are probably the most important. These consist of the methods used to pay for goods or services with a mobile handheld device, such as a smart cellular phone or Internet-enabled PDA. Dominant corporations are competing for the advance of their own standards, which will contribute to competition with their rivals. Among

different standards, the common issues are security, interoperability and usability. Mobile commerce security and payment are still in the adolescent stage. Many new protocols and methods are waiting to be discovered and developed.

REFERENCES

Ashley, P., Hinton, H., & Vandenwauver, M. (2001). Wired vs. wireless security: The Internet, WAP and iMode for e-commerce. In *Proceedings of Annual Computer Security Applications Conferences (ACSAC)*, New Orleans, Louisiana.

Biryukov, A., Shamir, A., & Wagner, D. (2000). Real time cryptanalysis of A5/1 on a PC. In *Proceedings of the 7th International Workshop on Fast Software Encryption*, New York City, New York.

Borisov, N., Goldberg, I., & Wagner, D. (2001). Intercepting mobile communications: The insecurity of 802.11. In *Proceedings of the 7th International Conference on Mobile Computing and Networking*, Rome, Italy.

Cam-Winget, N., Moore, T., Stanley, D., & Walker, J. (2002). IEEE 802.11i Overview. Retrieved July 1, 2004, from *http://csrc.nist.gov/wireless/S10_802.11i%20Overview-jw1.pdf*

ComputerWorld. (2000). Micorpayments. Retrieved July 1, 2004, from *www.computerworld.com/news/2000/story/0,11280,44623,00.html*

DeClercq, K. (2002). Banking sector, Lessius Hogeschool, Antwerp, Belgium.

Fluhrer, S., Martin, I., & Shamir, A. (2001). Weakness in the key scheduling algorithm of RC4. In *Proceedings of the 8th Annual Workshop on Selected Areas in Cryptography*, Toronto, Ontario, Canada.

ICSTIS (The Independent Committee for the Supervision of Standards of Telephone Information Services). (n.d.). Reverse-billed premium rate SMS. Retrieved February 17, 2004, from *www.icstis.org.uk/icstis2002/default.asp?node=6*

Jakobsson, M., & Wetzel, S. (2001). Security weaknesses in Bluetooth. *Topics in Cryptography: CT-RSA 2001.* LNCS 2020, 176-191. Berlin: Springer-Verlag.

Juul, N., & Jorgensen, N. (2002). Security issues in mobile commerce using WAP. In *Proceedings of the 15th Bled Electronic Commerce Conference*, Bled, Slovenia.

Ma, K., & Cao, X. (2003). Research of Bluetooth security manager. In *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing*, Nanjing, China.

Mobey Forum. (2002). Retrieved October 10, 2002, from *www.mobeyforum.org/*

Mobile electronic Transactions (MeT) Ltd. (2002). Retrieved November 22, 2002, from *www.mobiletransaction.org/*

Mobile Payment Forum. (2002). Retrieved December 15, 2002, from *www.mobilepaymentforum.org/*

Pahlavan, K., & Krishnamurthy, P. (2002). *Principles of Wireless Networks: A Unified Approach*. Upper Saddle River: Prentice Hall PTR.

Reuters. (2001). The Yankee Group publishes U.S. mobile commerce forecast. Retrieved October 16, 2003, from *http://about.reuters.com/newsreleases/art_31-10-2001_id765.asp*

Stubblefield, A., Ioannidis, J., & Rubin, A.D. (2002). Using the Fluhrer, Martin, and Shamir attack to break WEP. In *Proceedings of the Network and Distributed Systems Security Symposium*, San Diego, California.

Tanenbaum, A.S. (2002). *Computer networks* (4th ed.). Upper Saddle River: Prentice Hall PTR.

WAP (Wireless Application Protocol). (2003). Open Mobile Alliance Ltd. Retrieved November 21, 2002, from *www.wapforum.org/*

The Yankee Group. (2002). Over 50% of large U.S. enterprises plan to implement a wireless/mobile solution by 2003. Retrieved November 6, 2003, from *www.yankeegroup.com/public/news_releases/news_release_detail.jsp?ID=PressReleases/news_09102002_wmec.htm*

KEY TERMS

Micro/Macro Payment: A mobile payment of approximately $10 or less (often for mobile content such as video downloads or gaming) is called a micro payment, while a macro payment refers to a larger-value payment.

Mobile Commerce: The exchange or buying and selling of commodities, services or information on the Internet (wired or wireless) through the use of mobile handheld devices.

Mobile Commerce Security: The technological and managerial procedures applied to mobile commerce to provide security services for mobile commerce information and systems.

Mobile Payment: The transfer of financial value and corresponding services or items between different participants in mobile commerce systems.

Subscriber Identity Module (SIM): A device in the GSM that contains the subscriber's authentication information, such as cryptographic keys, and a unique identifier called international mobile subscriber identity (IMSI).

WAP Gap: A security weakness in WAP. It is caused by the inclusion of the WAP gateway in a security session such that encrypted messages sent by end systems might temporarily become clear text on the WAP gateway when messages are processed.

Wired Equivalent Privacy (WEP): The WEP is data link-level protocol that provides security for the IEEE 802.11 WLAN standards. The encryption algorithm used in WEP is a stream cipher based on RC4.

Mobile Computing for M–Commerce

Anastasis Sofokleous
Brunel University , UK

Marios C. Angelides
Brunel University, UK

Christos Schizas
University of Cyprus, Cyprus

INTRODUCTION

The ubiquitous nature of modern mobile computing has made "any information, any device, any network, anytime, anywhere" a well-known reality. Traditionally, mobile devices are smaller, and data transfer rates are much lower. However, mobile and wireless networks are becoming faster in terms of transfer rates, while mobile devices are becoming smaller, more compact, less power consuming, and, most importantly, user-friendly. As more new applications and services become available every day, the number of mobile device owners and users is increasing exponentially. Furthermore, content is targeted to user needs and preferences by making use of personal and location data. The user profile and location information is becoming increasingly a necessity.

The aim of this article is to present an overview of key mobile computing concepts, in particular, those of relevance to m-commerce. The following sections discuss the challenges of mobile computing and present issues on m-commerce. Finally, this article concludes with a discussion of future trends.

CHALLENGES OF MOBILE COMPUTING

Current mobile devices exhibit several constraints:

- Limited screen space: screens cannot be made physically bigger, as the devices must fit into hand or pocket to enable portability (Brewster & Cryer, 1999)
- Unfriendly user interfaces
- Limited resources (memory, processing power, energy power, tracking)
- Variable connectivity performance and reliability
- Constantly changing environment
- Security

These constraints call for immediate development of mobile devices that can accommodate high quality, user-friendly ubiquitous access to information, based on the needs and preferences of mobile users. It also is important that these systems must be flexible enough to support execution of new mobile services and applications based on a local and personal profile of the mobile user.

In order to evaluate the challenges that arise in mobile computing, we need to consider the relationships between mobility, portability, human ergonomics, and cost. While the mobility refers to the ability to move or be moved easily, portability relates to the ability to move user data along with the users. A portable device is small and lightweight, a fact that precludes the use of traditional hard-drive and keyboard designs. The small size and its inherent portability, as well as easy access to information are the greatest assets of mobile devices (Newcomb et al., 2003). Although mobile devices were initially used for calendar and contact management, wireless connectivity has led to new uses, such as user location tracking on-the-move. The ability to change locations while connected to the Internet increases the volatility of some information. As volatility increases, the cost-benefit trade of points shift, calling for appropriate modifications in the design.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

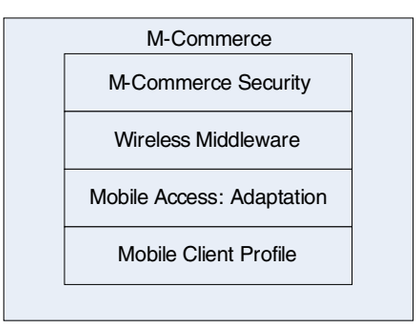
Wireless communications and mobile connectivity are overridden by bandwidth fluctuations, higher loss rates, more frequent and extended disconnections, and network failures that make Quality of Service (QoS) a continuous challenge. As a result, applications must adapt to a continuously changing QoS. Although mobile devices are designed to run light applications in a stand-alone mode, they still make use of wireless communication technologies such as Bluetooth, GPRS, and WiFi, which makes them useful in the new mobile world sphere, but they succumb to QoS limitations as a result of portability.

Mobility also is characterized by location transparency and dependency. A challenge for mobile computing is to factor out all the information intelligently and provide mechanisms to obtain configuration data appropriate to the current user location. In fact, in order to resolve a user's location, it is necessary to filter information through several layers: discovering the global position, translating the location, superimposing a map, identifying points of interest for the user and their relative range to that of the user. This suggests a multi-layer infrastructure. A number of location tracking services were developed in order to provide location information transparently to application developers who need to deploy location-aware applications.

M-COMMERCE

Mobile commerce is fast becoming the new trend for buying goods and services. As with e-commerce, it requires security for mobile transactions, middleware for content retrieval, and adaptation using client and device information.

Figure 1. M-commerce

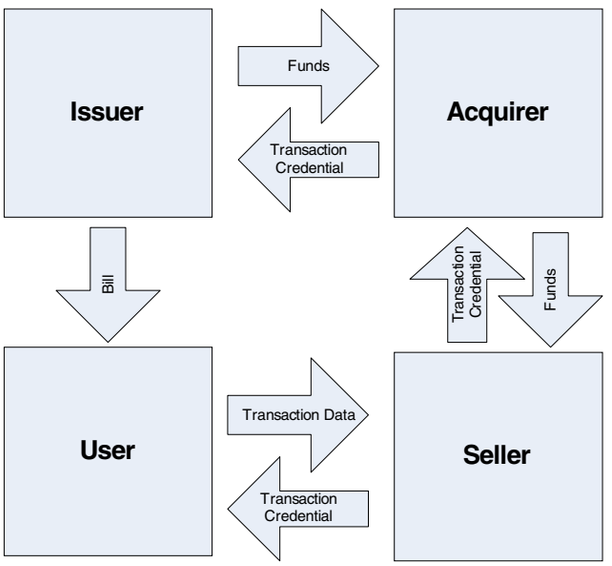


The enormous effect of mobile commerce in our lives can be noticed by studying the effect of m-commerce on industries in a way that will exceed wireline e-commerce as the method of preference for digital commerce transactions (e.g., financial services, mobile banking), telecommunications, retail and service, and information services (e.g., delivery of financial news and traffic updates). The global m-commerce market is likely to be worth a surprising US $200 billion in 2004 (More Magic Software, 2000). Report statistics confirm that in 2003, over a billion mobile phone users regarded it as a valuable communication tool. Global mobile commerce revenue projections show revenues up to the 88 billions for 2009 (Juniper Research, 2004).

Mobile security (M-Security) and mobile payment (M-Payment) are essential to mobile commerce and mobile world. Consumers and merchants have benefited from the virtual payments that information technology has conducted. Due to the extensive use of mobile devices nowadays, a number of payment methods have been deployed that allow the payment of services/goods from any mobile device. The success of mobile payments is contingent on the same factors that have fueled the growth of traditional non-cash payments: security, interoperability, privacy, global acceptance, and ease-of-use (Mobile Payment Forum, 2002).

The challenges associated with mobile payments are perhaps better understood using the example of credit card transaction. A card transaction involves at least four parties. As illustrated in Figure 2, the user as a buyer is billed by the card issuer for the goods and services he or she receives from the seller, and the funds are transferred from the issuer to the acquirer, and finally to the merchant. First, the consumer initializes the mobile purchase, registers with the payment provider, and authorizes the payment. A content provider or merchant sells product to the customer. The provider or merchant forwards the purchase requests to a payment service provider, relays authorization requests back to the customer, and is responsible for the delivery of the content. Another party in the payment procedure is the payment service provider, who is responsible for controlling the flow of transaction between mobile consumers, content providers, and trusted third parties (TTP), as well as for enabling and routing the payment message initiated from the mobile device to be cleared

Figure 2. A classic payment operation



by the TTP. A payment service provider could be a mobile operator, a bank, a credit card company, or an independent payment vendor.

Although with mobile payments, the payment transaction is similar to that described in Figure 1, there are some differences with regards to the transport of payment details, as this will involve a mobile network operator and will use either a browser-based protocol such as WAP or HTML or will be done via Bluetooth, WIFI, or infrared. The configuration of the payment mechanism could be achieved with the installation of either an applet or a specific application on a mobile device, and it usually takes place once. The first steps following successful installation include initialization of consumer payment (i.e., transferring payment information over a wireless network), user authentication, and payment completion, including receipt generation.

Existing mobile payment applications are categorized based on the payment settlement methods that they implement: pre-paid (using smart cards or digital wallet), instant paid (direct debiting or offline payments), and post paid (credit card or telephone bill) (Seema & Chang-Tien, 2004). Developers deploying applications using mobile payments must consider security, interoperability, and usability requirements. A secure application will allow an issuer to identify a user, authenticate a transaction, and prevent unauthorized parties from obtaining any information on the transaction. Interoperability guarantees completion of a transaction between different mobile devices or distribution of a transaction across devices. Usability ensures user-friendliness and multi-users.

M-COMMERCE SECURITY

Security is a crucial concern for anyone deploying mobile devices and applications, because personal information has to be delivered to a number of mobile workers engaged in online activities outside the secure perimeter of a corporate area. That increases the threat for unauthorized access and use of private and personal data. In order to authenticate the users accessing shared data, developers are using a number of authentication mechanisms, such as simple usernames and passwords, special single-use passwords from electronic tokens, cryptographic keys, and certificates from public key infrastructures (PKI). Additionally, developers are using authentication mechanisms to determine what data and applications the user can access (after login authorization). These mechanisms, often called policies or directories, are handled by databases that authenticate users and determine their permissions to access specific data simultaneously.

The current mobile business (M-Business) environment runs over the TCP/IPv4 protocol stack, which poses serious security level threats with respect to user authentication, integrity, and confidentiality. In a mobile environment, it is necessary to have identification and non-repudiation and service availability, mostly a concern for Internet and or application service providers. For these purposes, carriers (telecom operators and access providers), services, application providers, and users demand end-to-end security as far as possible (Leonidou et al., 2003; Tsaoussidis & Matta, 2002).

The technologies used in order to implement m-business services and applications like iMode, Handheld Device Mark-up Language (HDML) and Wireless Access Protocol (WAP) can secure the transport of data (encryption) between clients and servers, but they do not provide applicable security layers, espe-

cially user PIN-protected digital signatures, which are essential to secure transactions. Therefore, consumers cannot acknowledge transactions that are automatically generated by their mobile devices. Besides the characteristics of the individual mobile devices, some of the securities issues issued are dependent on the connectivity between the devices. Internet2 and IPv6 also have many security concerns, such as the authentication and authorization of binding updates sent from mobile nodes and the denial-of-service attack (Roe et al., 2002).

It is important to incorporate security controls when developing mobile applications rather than deploying the applications before and without fitting security. Fortunately, it is now becoming possible to implement security controls for mobile devices that do afford a reasonable level of protection in each of the four main problem areas: virus attacks, data storage, synchronization, and network security (Brettle, 2004).

WIRELESS MIDDLEWARE

Content delivery and transformation of applications to wireless devices without rewriting the application can be facilitated by wireless middleware. Additionally, a middleware framework can support multiple wireless device types and provide continuous access to content or services (Sofokleous et al., 2004). The main functionality of wireless middleware is the data transformation shaping a bridge from one programming language to another and, in a number of circumstances, is the manipulation of content in order to suit different device specifications. Wireless middleware components can detect and store device characteristics in a database and later optimize the wireless data output according to device attributes by using various data-compression algorithms, such as Huffman coding, Dynamic Huffman coding, Arithmetic coding, and Lempel-Ziv coding. Data compression algorithms serve to minimize the amount of data being sent over the wireless link, thus improving overall performance on a handheld device. Additionally, they ensure end-to-end security from handheld devices to application servers, and finally, they perform message storage and forwarding, should the user get disconnected from the network. They provide operation support by offering utilities and tools to allow MIS personnel to manage and troubleshoot wireless devices. Choosing the right wireless middleware is dependent on the following key factors: platform language, platform support and security, middleware integration with other products, synchronization, scalability, convergence, adaptability, and fault tolerance (Lutz,2002; Vichr & Malhotra, 2001).

MOBILE ACCESS ADAPTATION

In order to offer many different services to a growing variety of devices, providers must perform an extensive adaptation of both content (to meet the user's interests) and presentation (to meet the user device characteristics) (Gimson, 2002). The network topology and physical connections between hosts in the network must be constantly recomputed, and application software must adapt its behavior continuously in response to this changing context (Julien et al., 2003) either when server-usage is light, or if users pay for the privilege (Ghinea & Angelides, 2004).

The developed architecture of m-commerce communications exploits user perceptual tolerance to varying QoS in order to optimize network bandwidth and data sizing. This will provide QoS impacts upon the success of m-commerce applications without doubt, as it plays a pivotal role in attracting and retaining customers. As the content adaptation and, in general, the mobile access personalization concept are budding, central role plays the utilization of the mobile client profile, which is analyzed in the next section.

MOBILE CLIENT PROFILE

The main goal of profile management is to offer content targeted to users' needs and interests, using a presentation that matches their mobile device specification. Usually, this is done by collecting all the data that can be useful for identifying the content and the presentation that best fit the user's expectations and the device capabilities. The information may be combined with the location of the user and the action context of the user at the time of the request (Agostini et al., 2003).

In order to have a complete user profile, different entities are assembled from different logical locations (i.e., the personal data is provided by the user, whereas the information about the user's current location is usually provided by the network operator). Providers should query these entities to get the required information for a user. Several problems and methods for holding back the privacy of data are raised, as mobile devices allow the control of personal identifying information (Srivastava, 2004). People are instantly concerned about location privacy generated by location tracking services.

FUTURE TRENDS

During the past decade, computing and mobile computing have changed the business and consumer perception, and there is no doubt that mobile computing has already exceeded most expectations. Architectures and protocol standards, management, services, applications, and the human factor make possible the evolution of mobility (Angelides, 2004). The major areas that will be involved are the hardware, the middleware, the operating system and the applications.

In the area of software, while likely applications are being deployed, mobile services and applications will progressively distribute a variety of higher bandwidth applications, such as multimedia messaging, online gaming, and so forth. Several applications, such as transactional applications (financial services/banking, home shopping, instant messages, stock quotes, sale details, client information, location-based services, etc.) have already showed a tremendous potential for growth. Unfortunately, applications are restricted by the available hardware and software resources. As a result, portable devices must be robust, reliable, user-friendly, enchanting, functional, and expandable.

Additionally, mobile computing devices will have to provide a similar level of security and interoperability as usual handsets, combined with a performance level approaching that of desktop computers. The variety of wireless connectivity solutions, the operating systems, the presentation technologies, the processors, the battery technologies, the memory options, and the user interfaces are analyzed and examined, as they will enable the growth of mobile computing. The operating system also is largely dependent on the hardware, but it should be scalable, customizable, and expandable.

Third generation mobile communication systems, such as Universal Mobile Telecommunications System (UMTS), will provide optimum wireless transmission speeds up to 2 Mbits/s, and they will have voice and video connections to the mobile devices.

The future of mobile computing is looking very promising, and as wireless computing technology is being gradually deployed, the working lifestyle may change, as well.

Figure 3. Areas of mobility evolution

CONCLUSION

This article discusses the more important issues that affect m-commerce. The ability to access information on demand while mobile will be very significant. IT groups need to understand the ways mobile and wireless technology could benefit m-commerce and avoid deploying wireless on top of wired, which adds incremental costs. Mobile application frameworks create a range of new security exposures, which have to be understood and taken under consideration during the design steps of the mobile frameworks. In the general view, e-commerce is concerned with trading of goods and services over the Web and the m-commerce with business transactions conducted while on the move. However, the essential difference between e-commerce and m-commerce is neither the wire nor the wireless aspects, but the potential to explore opportunities from a different perspective.

Companies need to customize the content in order to meet the requirements imposed by bandwidth and the small display size of mobile devices. Mobile services and applications, such as location management, locations, profile-based services, and banking services are some of the applications that have great potential for expansion. The demand of m-business applications and services will grow, as new developments in mobile technology unfold. Nowadays, challenging mobile payment solutions have already estab-

lished their position in the marketplace. As software systems are becoming more complex and need to extend to become wireless, in some instances, it may be useful to use a wireless middleware. What we are currently observing is mobile computing becoming increasingly pervasive among businesses and consumers.

REFERENCES

Agostini, A., Bettini, C., Cesa-Bianchi, N., Maggiorini, D., & Riboni, D. (2003). Integrated profile management for mobile computing. *Proceedings of the Workshop on Artificial Intelligence, Information Access, and Mobile Computing,* Acapulco, Mexico.

Angelides, M.C. (2004). Mobile multimedia and communications and m-commerce. *Multimedia Tools and Applications, 22*(2), 115-116.

Brettle, P. (2004). White paper on mobile security. Insight consulting. Retrieved from *http://www.insight.co.uk*

Brewster, A.S., & Cryer, P.G. (1999). Maximizing screen-space on mobile computing devices. *Proceedings of the Conference on Human Factors in Computing Systems,* Pittsburgh. New York.

Dahleberg, T., & Tuunainen, V. (2001). Mobile payments: The trust perspective. *Proceedings of the International Workshop Seamless Mobility,* Sollentuna, Spain.

Ghinea, G., & Angelides, C.M. (2004). A user perspective of quality of service in m-commerce. *Multimedia tools and applications, 22*(2), 187-206.

Gimson, R. (2002). Delivery context overview for device independence [W3C working draft]. Retrieved September 12, 2002 from *http://www.w3.org/2001/di/public/dco/dco-draft-20020912/*

Julien, C., Roman, G., & Huang, Q. (2003). Declarative and dynamic context specification supporting mobile computing in ad hoc networks [Technical Report WUCSE-03-13]. St. Louis, MO, Washington University.

Juniper Research. (2004). The big micropayment opportunity [White paper]. Retrieved September 24, 2002 from *http://industries.bnet.com/abstract.aspx?seid=2552&docid=121277*

Leonidou, C., et al. (2003). A security tunnel for conducting mobile business over the TCP protocol. *Proceedings of the 2nd International Conference on Mobile Business,* Vienna, Austria.

Lutz, E.W. (2002). Middleware for the wireless Web. Faulkner Information Services. Retrieved August 25, 2004 from *http://www.faulkner.com*

Mobile Payment Forum. (2002). Enabling secure, interoperable, and user-friendly mobile payments. Retrieved August 18, 2004 from *http://www.mobilepaymentforum.org/pdfs/mpf_whitepaper.pdf*

More Magic Software. (2000). Payment transaction platform. Retrieved July 25, 2003 from *http://www.moremagic.com/whitepapers/technical_wp_twp021c.html*

Newcomb, E., Pashley, T., & Stasko, J. (2003). Mobile computing in the retail arena. *ACM Proceedings of the Conference on Human Factors in Computing Systems,* Florida.

Roe, M., Aura, T., & Shea, G.O. (2002). Authentication of Mobile IPv6 Binding Updates and Acknowledgements. (Internet draft). Retrieved August 10, 2004 from *http://research.microsoft.com/users/mroe/cam-v3.pdf*

Seema, N., & Chang-Tien, L. (2004). *Advances in security and payment methods for mobile commerce*. Hershey, PA: Idea Group Publishing.

Sofokleous, A., Mavromoustakos, S., Andreou, A.S., Papadopoulos, A.G., & Samaras, G. (2004). JiniusLink: A distributed architecture for mobile services based on localization and personalization. *Proceedings of the IADIS International Conference,* Lisbon, Portugal.

Srivastava, L. (2004). Social and human consideration for a mobile world. *Proceedings of the ITU/MIC Workshop on Shaping the Future Mobile Information Society,* Seoul, Korea.

Tsaoussidis, V., & Matta, I. (2002). Open issues on TCP for mobile computing. *Journal of Wireless Communications and Mobile Computing*, *2*(1).

Vichr, R., & Malhotra, V. (2001). Middleware smoothes the bumpy road to wireless integration. *IBM.* Retrieved August 11, 2004 from *http://www-106.ibm.com/developerworks/library/wi-midarch/index.html*

KEY TERMS

E-Commerce: The conduct of commerce in goods and services over the Internet.

Localization: The process to adapt content to specific users in specific locations.

M-Business: Mobile business means using any mobile device to make business practice more efficient, easier, and profitable.

M-Commerce: Mobile commerce is the transactions of goods and services through wireless handheld devices, such as cellular telephones and personal digital assistants (PDAs).

Mobile Computing: Mobile computing encompasses a number of technologies and devices, such as wireless LANs, notebook computers, cell and smart phones, tablet PCs, and PDAs, helping the organization of our life, communication with coworkers or friends, or the accomplishment of our jobs more efficiently.

Mobile Device: Mobile device is a wireless communication tool, including mobile phones, PDAs, wireless tablets, and mobile computers (Mobile Payment Forum, 2002).

Mobility: The ability to move or to be moved easily from one place to another.

M-Payment: Mobile payment is defined as the process of two parties exchanging financial value using a mobile device in return for goods or services (Seema Nambiar, paper).

M-Security: Mobile security is the technologies and methods used for securing wireless communication between the mobile device and the other point of communication, such as another mobile client or a pc.

Profile: Profile is any information that can be used to offer a better response to a request (i.e., the information that characterizes the user, the device, the infrastructure, the context, and the content involved in a service request) (Agostini et al.,2003).

Wifi: Wifi (wireless fidelity) is a technology that covers certain types of wireless local area networks (WLANs), enabling users to connect wirelessly to a system or wired local network and use specifications in the 802.11 family.

Mobile Location Based Services

Bardo Fraunholz
Deakin University, Australia

Jürgen Jung
Uni Duisburg-Essen, Germany

Chandana Unnithan
Deakin University, Australia

INTRODUCTION

Mobility has become a key factor around the world, as the use of ubiquitous devices, including laptops, personal digital assistants (PDAs), and mobile phones, are increasingly becoming part of daily life (Steinfield, 2004). Adding mobility to computing power, and with advanced personalization of technologies, new business applications are emerging in the area of mobile communications (Jagoe, 2003). The fastest growing segment among these applications is location-based services. This article offers a brief overview of services and their supporting technologies, and provides an outlook for their future.

BACKGROUND

The popularity and usage of mobile devices and communications are on the rise, due to convenience as well as progress in technology. This section initially takes a closer look at the underlying statistics and then tries to define these services from a synopsized view of many authors.

While industrialized nations have imbibed mobile technologies by almost transitioning technologies, even in developing nations, mobile communication has taken over fixed-line services (ITU, 2003). This progress is driven by mobile network operators who continue to look for potential revenue-generating business models in order to increase the demand for services, as there is increased competition reducing prices for voice services. One of the popular and progressive business models is mobile location-based services for the Global Systems for Mobile communications (GSM) networks. These services provide customers with a possibility to get information, based on their location. Such information may be, for example, the nearest gas station, hotel, or any similar service that might be stored by the service provider, in relation to any particular locality. These services are location-aware applications (VanderMeer, 2001) that take the user's location into account in order to deliver a service.

Location-based applications have developed into a substantial business case for mobile network operators during the last few years (Steinfield, 2004). ITU estimates worldwide revenues from LBS would exceed US $2.6 billion in 2005 and reach US $9.9 billion by 2010 (Leite & Pereira, 2001). Market research by Strategy Analytics in 2001 indicated that these services have a revenue potential of US $6 billion of revenue in Western Europe and US $4.6 billion in North America by the end of 2005 (Paavalainen, 2001). An ARC Group study predicted that these services will account for more than 40% of mobile data revenues worldwide by 2007 (Greenspan, 2002). According to Smith (2000), more than half of the US mobile customer base was willing to accept some form of advertising on a mobile handset, if they were able to use location services for free. An Ovum study predicts Western European market to touch US $6.6 billion by 2006 (Greenspan, 2002).

Mobile subscribers, especially in industrialized societies, are unwittingly using a location determination technology (Steinfield, 2004) due to the fact that regulators in most of these nations have initiated rules requiring network operators to deliver information about the location of a subscriber to public safety answering points in the event of an emergency. In the US, the Federal Communications Commission requires operators to provide the location of all mobile

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

emergency calls, and, therefore, the market itself was government driven (FCC, 2003). The European Union is developing a similar requirement for its emergency services (D'Roza & Bilchev, 2003). Corporations have begun to realize the benefit in deploying these cost-effective services in order to increase the efficiency of field staff (Schiller, 2003).

Prasad (2003) purports that location-based service is the ability to find the geographical location of the mobile device and provide services based on this location information. Magon & Shukla (2003) agree that that it is the capability to find the geographical location of the mobile device and then provide services based on this location information. The concept of these services is based on the ability to find the geographical location of the mobile device and provide services based on this location information. Therefore, they can be described as applications, which react according to a geographic trigger. A geographic trigger might be the input of a town name, zip code, or street into a Web page, the position of a mobile phone user, or the precise position of your car as you are driving home from the office (whereonearth, 2003).

In the popular context, mobile location services have become solutions that leverage positional and spatial analysis tools (location information) to deliver consumer applications on a mobile device (Jagoe, 2003). Currently, these services are at the conjuncture of geographic information systems and the wireless networking industries. Location information analysis technologies developed for Geographic Information Systems have been repurposed for the speed and scalability of mobile location-based services. Positioning technologies leverage wireless and satellite technologies to perform complex measurements to pinpoint the location of a mobile user—a critical piece of information in mobile location-based applications. Mobile data networks are used for application deployment. The following section aims at characterizing positioning technologies that support mobile location-based services.

CHARACTERIZING POSITIONING TECHNOLOGIES

The critical factor for mobile location-based service is the determination of a user's location, using positioning technologies. Drane and Rizos (1998) emphasize three conceptually different approaches to generic positioning technologies, such as signpost, wave-based systems, and dead reckoning. Within the mobile communication networks, Röttger-Gerigk (2002) distinguishes between network-based and specialized positioning services.

Sign-post systems represent the simplest sort of positioning, which is based on an infrastructure of signposts (i.e., landmarks or beacons). Positions are measured by determining the nearest beacon to the mobile object. Therefore, positioning is reduced to the statement that a mobile object is nearby or in certain proximity of a certain beacon. The accuracy of signpost systems is given by the distance between two neighboring signposts. Currently, signpost systems are used for automatic toll collection on highways (Hills & Blythe, 1994).

Wave-based positioning systems use propagation properties of usually electro-magnetic waves to determine the position of a mobile object. Locations of mobile objects are determined relative to one ore more reference sites. The availability of wave-based positioning systems is limited by an undisturbed reception of the radio waves sent by the reference points.

Dead reckoning systems consist of several vehicle-mounted sensors for the detection of a mobile object's movements. These sensors are used for the continuous determination of a vehicle's velocity and heading. Starting from an initial reference point, a mobile object can be located by logging its speed and heading over time.

Another classification of positioning technologies uses the approach as to where the location of a mobile object is determined (Röttger-Gerigk, 2002). Here, positioning systems are characterized as self-positioning or remote positioning. In self-positioning systems, the position is determined in the mobile device itself. Hence, the position is primarily known by the mobile object itself. Complementary, the information about the location may be transmitted to external systems or partners over a mobile communication infrastructure. Remote positioning systems provide positioning services only for external systems, which can then use this information for customized location base.

The hitherto presented types of positioning technologies usually result in an absolute specification of a mobile user's location. Signpost systems specify a position based on a network of landmarks and wave-

based systems on the basis of properties of the propagation of electro-magnetic waves. Dead reckoning systems record movements, acceleration, and the velocity of mobile objects by using special sensors. Nevertheless, mobile users (especially the ones going by car) are moving along roads. The exact determination of a mobile user's position is supported by its estimated position in relation to given map data (Drane & Rizos, 1998). One heuristic might be that a user in a car might only drive on a given road.

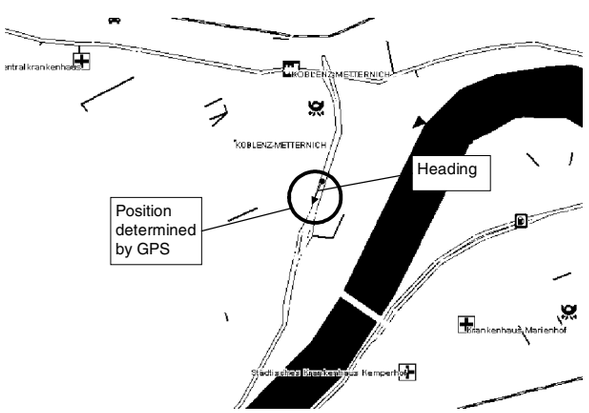
One example of the combination of established positioning services and map matching is shown in Figure 1. The estimated position of the mobile user is given by the circle in the diagram. According to the simple rule that a mobile user in a car can only be located on a road results in the positioning of this user on the given position in the figure.

Practically, several of the given positioning services are combined. The result is a high-value positioning service. Popular navigation systems, for example, depend on GPS, dead reckoning, and map matching. A GPS antenna is used for the determination of a vehicle's position, and this information is adjusted with the information given by dead reckoning and map matching. Hence, different positioning systems cannot be discussed in an isolated manner. Current systems basically depend on basic kinds of positioning technologies as well as valuable combinations of those technologies.

Global Positioning System

The basic conceptualizations of special positioning technologies (like GPS) and different kinds of network-based positioning services are relevant to mobile location-based services. The Global Positioning System (GPS) is a self-positioning, wave-based positioning system launched by the U.S. Department of Defense in the 1970s (Drane & Rizos 1998). Currently, GPS consists of at least 24 satellites revolving around the earth on six orbits (Lechner & Baumann, 1999). All satellites send a continuous radio signal every second, including its position and the sending time. A special GPS receiver uses the signals of at least three satellites for the determination of its global position. The position is computed by propagation delays of the signals sent by the satellites. Similar but less popular systems are the Russian GLONASS and the future European satellite-based positioning system GALILEO. With respect to accuracy, satellite-based positioning systems are expected to play an important role in the long term.

Figure 1. Map matching in positioning services



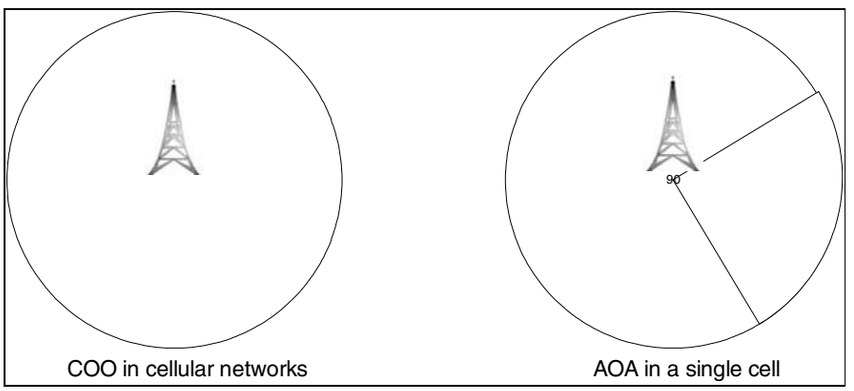
Network-Based Positioning

Network-based positioning is usually part of another given network. Examples of such kinds of networks are cellular communication networks such as GSM (Global System for Mobile telecommunication) and UMTS (Universal Mobile Telecommunication System) (Röttger-Gerigk, 2002; Steinfield, 2004). This positioning technology is currently most relevant to mobile location-based services.

Cell of Origin (COO) determines a mobile user's location by the identification of the cell in which the person's mobile device is registered. Hence, the accuracy of COO is given by the size of a cell. This positioning method is also known as Cell Global Identity (CGI). Despite its comparatively low accuracy, this technology is widely used in cellular networks. The reasons are simple: the accuracy is sufficient for some applications, and the service is implemented in all GSM-based networks. COO is a remote positioning service (like most network-based positioning services), but information about a location also can be transferred to the mobile device by cell broadcast.

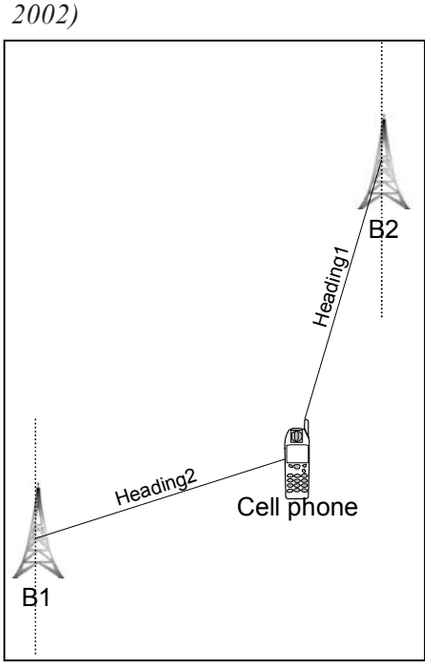
Angle of Arrival (AOA) is based on traditional positioning techniques and uses the bearing of at least two base stations. In most cellular networks, such as GSM, the antennas of a base station might be used for the determination of the angle of an incoming

Figure 2. Positioning by cell ID (left) and arc of a circle



signal. The antenna of a base station in GSM only covers a part of the area of a circle (i.e., 120° of a whole circle). Figure 2 illustrates the difference between COO and AOA: COO covers the whole of a network cell, whereas AOA only covers the arc of a circle. AOA is like COO, available in most cellular networks and, thus, already implemented. Using a single base station, the positioning accuracy is better than using COO and can be improved by combining the information of at least two base stations. Using the bearing of two base stations is displayed in Figure 3. Each of the base stations, B1 and B2, in Figure 3 receives the signal sent by a cell phone from a different angle (represented by headings 1 and 2).

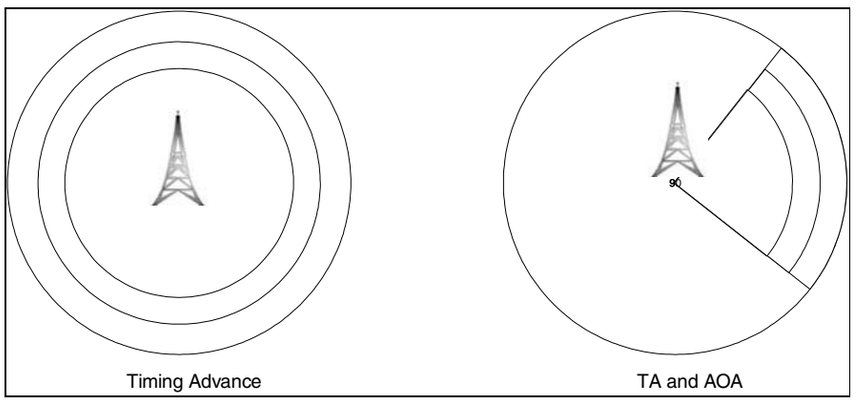
Figure 3. AOA with two base stations (Röttger-Gerigk, 2002)



Timing Advance (TA) is a very important function in GSM, because a time-multiplexing transmission method is used. Every data package has to fit into a given time slot. Because of the light-speed, the radio signal sent by a mobile device needs some time to reach the base station. Such a delay of a data packet has to be taken into account. TA determines the signal's running time and causes the mobile device to send the data some microseconds in advance. The timing advance allows for the determination of the distance between a base station and a mobile device in multiplies of 550m. Positioning using TA is shown in Figure 4. The diagram on the left side illustrates TA in a single cell, and the one on the right side combines TA and AOA. TA is actually a GSM-specific method for the determination of the distance between a base-station and a mobile device. Nevertheless, it demonstrates the basic idea of distance measurement in cellular networks. TA is not only a hypothetical method, but also is practically used in GSM. Similar methods are used in other cellular networks.

In the Time Difference of Arrival (TDOA) method, the time difference of the arrival of a signal sent by one single mobile device at several (at least three) base stations is recorded. In other words, a mobile unit sends a specific signal at a given time. This signal is received by several base stations at a later moment. The expansion speed (light speed) and the time differences of arrival at the base stations allow the positioning of the mobile unit. Essential for this positioning service are a precise time basis and a central unit (called Mobile Location Center) for the

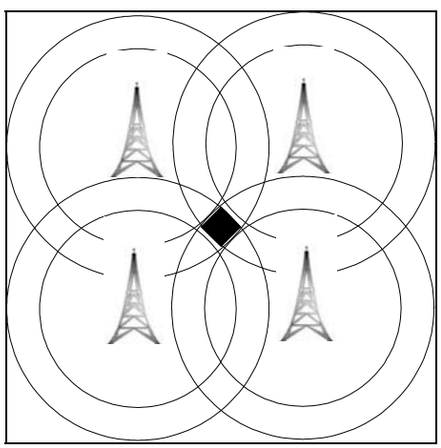
Figure 4. Positioning based on timing advance



synchronization of time data between base stations. TDOA is a remote-positioning service that needs no upgrade at the mobile unit and minor changes at the net infrastructure.

Time of Arrival (TOA) is a method similar to TDOA. In contrast to TDOA, the running time of a radio signal will be measured, and not the time. The mobile unit is sending a signal that will be received by at least three base stations. The position of the mobile device will be calculated on the basis of time differences of the received signal at each base station. A schematic drawing is given in Figure 5. Four base stations are used for the positioning of a mobile user. This user can be located in discrete distances from these four base stations. The distances of the user from the base stations, combined with the absolute positions of the base stations, allow the absolute localization of the mobile unit.

Figure 5. TOA with four base stations



Comparison of Positioning Technologies

The differences between GPS and network-based positioning systems are listed in Table 1.

All positioning systems depend on conceptual strengths and weaknesses. GPS profits from high accuracy in conjunction with high user device costs. Nevertheless, the positioning quality of GPS-based systems is hindered by a reduced reception of the satellite signal on certain roads. High buildings and trees may keep the GPS reception from the receiver. In current navigation systems, such differences from the GPS signal are compensated by dead reckoning systems. Combining GPS position signals with the data from in-vehicle velocity and direction sensors leads to a more precise positioning quality. The most obvious technology behind mobile location-based services is the positioning technologies and the widely recognized Global Positioning System (GPS). However, there are network-based positioning technologies that typically rely on triangulation of a signal from cell sites serving a mobile phone, and the serving cell site can be used as a fix for locating the user (Mobilein, 2003). There is a need to support multiple location determination technologies (LDT) and applications for locating the mobile device. An integrated solution should support many types of available LDT technologies, such as Cell ID, AOA, TDOA, GPS, and TOA (Infoinsight, 2002). The next section takes a closer look at some business examples of mobile location-based services.

Table 1. Comparison of GPS and network-based systems

GPS	Network
Network of its own	Part of a popular network
Special end-user devices	Widespread end user devices
High accuracy	Lower accuracy
Global availability	Availability restricted by network coverage

BUSINESS APPLICATIONS

Location-based services are added value services that depend on a mobile user's geographic position (Infoinsight, 2002). There are numerous ways in which location-based data can be exploited, especially in combination with user profiles, to offer solutions to customers (Steinfield, 2004). Pull services are requested by users once their location is determined, and push services are triggered automatically once a certain condition is met, such as crossing a boundary (D'Roza & Bilchev, 2003). Many of the services are offered by network operators or as value-added services with other organizations. Some of the network operator-based services include location-based information provision (Mobilein, 2003), location-sensitive billing (InfoInsight, 2002), entertainment, communication, transactions, and proximity services (Levijoki, 2001), mobile office, and business support services (Van de Kar & Bowman, 2001). Mobile network operators and other organizations, including health care/insurance providers, hotels, automobile companies, and so forth, work together to provide location-based services such as emergency and safety services, roadside assistance, travel information, traffic monitoring, and so forth (InfoInsight, 2002; Mobilein, 2003).

NEXTBUS in San Francisco uses an Internet-enabled mobile phone, or PDA, where bus riders can find estimated arrival times at each stop in real time, and also location-based advertisements will pop up on your mobile (Turban et al., 2002) (e.g., you have the time to get a cup of coffee before the bus arrives, and Starbuck's is 200 feet to the right). Hotelguide.com stores user profiles, specifically business travelers. At a new location, the user is able to search for a suitable hotel using the WAP phone, make a reservation, and book a taxi to get them to the hotel. Travelers in unfamiliar cities, who need immediate accommodations, find this business model very useful.

One of the largest computerized travel reservation systems (Galileo) offers a service to enable travelers to rebook and monitor the status of flights using WAP phones (Steinfield, 2004). They have the provision to notify the customer if the flights are delayed or canceled. In the US, the Federal Communications Commission issued the E911 mandate, requiring every network operator to be able to detect the location of subscribers within 50 meters for 67% of emergency calls, and 150 meters for 95% of calls (FCC, 2003). Dialing 911 from a mobile phone pinpoints your location and relays it to appropriate authorities, and the FCC mandates a degree of accuracy in the pinpointing for all mobile users in the US. The European Union has developed similar requirements for their E112 emergency services. Proximity services inform users when they are within a certain distance from others, businesses, and so forth. NTT DoCoMo offers a friends finder service on iMode, where you can find a predefined friend's location.

GM's Onstar is using vehicle-based GPS receivers and mapping/route guide services in selected cars. These services can be integrated with real-time traffic data to make routes contingent on traffic conditions. GPS North America (Gpsnorthamerica, 2003) has a Web application called MARCUS, which has the ability to locate and find a single vehicle, a fleet of vehicles, and the closest unit to a particular location address. This is updated every five minutes and can be seen in real-time as well as historical track or breadcrumb trial in the past three months. This application is designed for occurrences to allow remote monitoring of the fleet and crew. Automatic vehicle location in transit is another application that is growing and is expected to benefit in increased overall dispatching and operating efficiency, as well as more reliable service, as the system operates by measuring the real-time position of each vehicle.

Another useful service is used by some organizations in partnership with network operators (Crisp, 2003). Field staff is given access to internal data-

base systems on a continued basis and provided with a PDA. Take the scenario where the employee is in close proximity to a client, and the internal information database suggests critical updates to the client details. Relevant information may be passed on to the employee direct to the PDA. A similar application is rescheduling employee tasks in the field, taking into their current location. An employee may be able to finish work early and may be able to take another client call. If the organization is able to track the location of the employee, it is possible to reschedule the work using a PDA or mobile phone.

FUTURE TRENDS

Two concepts that are emerging are location awareness and sensitivity (Kleiman, 2003). Location awareness refers to applications or services that make use of location information, where location need not be the primary purpose of the application or service. In contrast, location sensitivity refers to location-enabled devices such as mobile phones, PDAs, or pagers. In the future, the phone will be able to locate a person where that person is and search for a suitable hotel without the need for the person entering the search. With third generation mobile technologies, the ability to track people wherever they are and notify customers of canceled flights in advance should become reality. Governments are moving to require that mobile operators develop the capability to automatically identify subscriber location so that in the event of emergency, the data may be forwarded to the public safety answering point to coordinate dispatch of emergency personnel. Combined with telemedicine techniques that allow psychological data transmission back to health care providers, this is another useful application. With the provision of 3G mobile technologies, it also may be possible to trace the person automatically, without the need for dialing emergency services (i.e., 911 in the US), being a context-aware, always-on technology.

CONCLUSION

Mobile location-based services is a confusing array of changing requirements, emerging standards, and rapidly developing technologies. There seems to be an unpredictable confluence of previously independent technologies, as each technology develops at a different rate, per the demands of its market, while being constrained by standards specifications. Many different players are involved in mobile location-based services, including mobile network operators, content providers, handset manufacturers, organizations, and so forth. Since all are stakeholders who potentially earn revenue from mobile location-based services, they require standard formats and interfaces to work efficiently. Otherwise, the costs of launching each service would be passed on to end users, and that would be destructive for mobile operators. The global third generation partnership project (3GPP), through which various standard bodies are attempting to create a smooth transition to third generation wireless networks, deals with mobile location-based services.

REFERENCES

Adams, P., Ashwell, G., & Baxter, R. (2003). Location based services—An overview of standards. *BT Technology Journal, 21*(1), 34-43.

Chatterjee, A. (2003). Role of GPS navigation, fleet management and other location based services. Retrieved December 11, 2003, from *http://www.gisdevelopment.net/technology/gps/techgp0045pf.htm*

Crisp, N. (2003). Open location based services, an Intelliware report. Retrieved November 12, 2003, from *www.intelliware.com*

D'Roza, T., & Bilchev, G. (2003). An overview of location based services. *BT Technology Journal, 21*(1), 20-27.

Drane, C., & Rizos, C. (1998). *Positioning systems in intelligent transportation systems*. Boston: Artech House.

FCC. (2003). Enhanced 911, Federal Communications Commission. Retrieved December 11, 2003, from *http://www.fcc.gov/911/enhanced/*

Gpsnorthamerica. (2003). How GPS North America works for you. *GPSNorthAmerica.co*m, Retrieved November 12, 2003, from *http://www.gpsnorthamerica.com/how.htm?trackcode=bizcom*

Greenspan, R. (2002). Locating wireless revenue, value. *CyberAtlas Wireless Markets*. Retrieved December 17, 2003, from *http://cyberatlas.internet.com/markets/wireless/article/0,,10094_1454791,00.html*

Hills, P., & Blythe, P. (1994). Automatic toll collection for pricing the use of road space—Using microwave communications technology. In I. Catling (Ed.), *Advanced technology for road transport* (pp. 119-144). Boston: Artech House.

Infoinsight. (2002). What are location services? *Info Insight*. Retrieved December 11, 2003, from *http://www.infoinsight.co.uk/etsi.htm*

ITU. (2003). ICT free statistics. *International Telecommunication Union.* Retrieved December 11, 2003, from *http://www.itu.int/ITU-D/ict/statistics/*

Jagoe, A. (2003). *Mobile location services—The definitive guide*. NJ: Pearson Education.

Kleiman, E. (2003). Combining wireless location services with enterprise ebusiness applications. Retrieved December 11, 2003, from *http://www.gisdevelopment.net/technology/lbs/techlbs007pf.htm*

Lechner, W., & Baumann, S. (1999). Grundlagen der Verkehrstelematik. In H. Evers, & G. Kasties (Eds.), *Kompendium der verkehrstelematik—Technologien, applikationen, perspektiven* (pp. 143-160). Köln,Germany: TÜV-Verlag.

Levijoki, S. (2001). *Privacy vs. location awareness* [unpublished]. Helsinki: Helsinki University of Technology.

Mobilein. (2003). Location based services. *Mobile in a minute*. Retrieved December 11, 2003, from *http://www.mobilein.com/location_based_services.htm*

Paavalainen, J. (2001). *Mobile business strategies, wireless press*, London: Addison-Wesley.

Prasad, M. (2003). Location based services. Retrieved December 11, 2003, from *http://www.gisdevelopment.net/technology/lbs/techlbs003pf.htm*

Röttger-Gerigk, S. (2002). Lokalisierungsmethoden. In W. Gora, & S. Röttger-Gerigk (Eds.), *Handbuch mobile-commerce* (pp. 419-426). Berlin: Springer.

Schiller. (2004). *Mobile communications*. London: Addison-Wesley.

Searby, S. (2003). Personalisation—An overview of its use and potential. *BT Technology Journal, 21*(1), 13-19.

Steinfield, C. (2004). The development of location based services in mobile commerce. In B. Preissl, H. Bouwman, & C. Steinfield (Eds.), *Elife after the dot.com bust* (pp. 177-197). Berlin: Springer.

Turban, E., King, D., Lee, J., Warkentin, M., & Chung, H.M. (2002). *Electronic commerce—A managerial perspective*. New Jersey: Pearson Education International.

Van de Kar, E., & Bowman, H. (2001). The development of location based mobile services. *Proceedings of the Edispuut Conference*, Amsterdam.

Whereonearth. (2003). What are location based services? *Whereonearth.* Retrieved December 11, 2003, from *//http://www.whereonearth.com/lbs*

KEY TERMS

AOA or Angle of Arrival: A positioning technique that determines a mobile user's location by the angle of an incoming signal. AOA covers only the arc of a circle instead of the whole cell.

COO Cell or Origin: A positioning technique that determines a mobile user's location by identifying a cell in which the person's mobile device is registered. Also known as Cell Global Identity (CGI).

GIS or Geographical Information Systems: Provide tools to provision and administer base map data such as built structures (streets and buildings) and terrain (mountain, rivers, etc.).

GPS or Global Positioning System: A self-positioning, wave-based positioning system consisting of 24 satellites revolving around the earth in six orbits, which send continuos radio signals using triangulation to determine an exact location.

GSM or Global System: For mobile telecommunications, a digital cellular communication network standard.

LDT: Location determination technologies.

LM or The Location Manager: A gateway that aggregates the location estimates for the mobile device from the various LDTs, computes the user location, and estimates the certainty of that location before being forwarded to the application.

Mobile Location Based Services: Applications that leverage positioning technologies and location information tools to deliver consumer applications on a mobile device.

PDA or Personal Digital Assistant: Refers to any small hand-held device that provides computing and data storage abilities.

TA or Timing Advance: A GSM-specific method for determining the distance between a base station and a mobile device.

TDOA or Time Difference of Arrival: A remote positioning service where the time difference of arrival of a radio signal sent by one single mobile device at several base stations is recorded.

TOA or Time of Arrival: Similar remote positioning service to TDOA, where the running time of a radio signal is measured and not the time.

UMTS or Universal Mobile Telecommunication System: A 3rd generation mobile communication network standard.

Mobile Multimedia for Commerce

P.M. Melliar-Smith
University of California, Santa Barbara, USA

L.E. Moser
University of California, Santa Barbara, USA

INTRODUCTION

The ready availability of mobile multimedia computing and communication devices is driving their use in commercial transactions. Mobile devices are lightweight and wireless so users can carry them and move about freely. Such devices include cell phones, PDAs and PCs equipped with cellular modems.

In the history of man, mobile commerce was the conventional form of commerce but during the twentieth century, it was superseded by fixed locations as a result of non-mobile infrastructure (stores and offices) and the ability of customers to travel. With modern mobile infrastructure, commerce can be conducted wherever the customer is located, and the sales activity can occur wherever and whenever it is convenient for the customer.

BACKGROUND

Mobile computing and communication devices, based on cellular communication, are a relatively recent innovation. Multimedia computing and communication, including video, audio, and text, are available for mobile devices but are limited by small screens, low bandwidth, and high transmission costs. These limitations distinguish mobile multimedia computing and communication from desktop multimedia computing and communication over the Internet, including WiFi, and dictate a somewhat different approach.

Mobile commercial processes are still largely experimental and are not yet well established in practice. Some researchers (Varshney, 2000) have projected that the use of mobile devices in consumer-to-business transactions will increase as much as 40%. Cautious consumers, inadequate mobile devices, security concerns, and undeveloped business models and procedures currently limit the use of mobile multimedia devices for commercial transactions.

Because mobile multimedia commerce using mobile devices is a new and developing field, there is relatively little available information, and that information is scattered. Early discussions of mobile commerce can be found in Senn (2000) and Varshney (2000). The i-mode service (Kinoshita, 2002; Lane, 2002) for mobile commerce has achieved some commercial success, within the limitations of existing devices and protocols.

LIMITATIONS OF THE MOBILE DEVICE

Cellular communication is wireless communication between mobile devices (e.g., cell phones, PDAs, and PCs) and fixed base stations. A base station serves mobile devices within a relatively small area of a few square miles (called a cell). The base stations are interconnected by fixed telecommunication infrastructure that provides connection with other telecommunication systems. When a mobile device passes from the cell of one base station to that of another, the first base station hands off communication with the device to the other, without disrupting communication.

Mobile devices are inherently more limited than fixed devices, but these limitations, appropriately recognized and accommodated, do not preclude their use in commerce (Buranatrived, 2002; Lee & Benbasat, 2003). Mobile devices have restricted display, input, print, and communication capabilities. The impact of these limitations depends on the user. A professional mobile sales representative needs better display, input, and print capabilities than many other kinds of users.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Mobile devices, such as cell phones and PDAs, have very small displays (less than 15 cm) that are likely to remain small, a limitation imposed by the need to insert the device into a pocket or purse, or to carry the device on a belt, and also by battery consumption. Such displays are inadequate for viewing detailed textual or graphical material. In an environment that is saturated with television, video, animated Web pages, and so forth, impressive multimedia sales presentations are even more important. Therefore, a mobile sales representative most likely will carry a notebook computer with a high-resolution display of 30 cm to 50 cm, and might even carry a projection display, which imposes little limitation on the material to be displayed.

The input capabilities of current mobile devices, such as cell phones and PDAs, are currently primitive and difficult to use for commercial activities. When natural language voice input is improved, the input of more complex requests, responses, and textual material will be possible. Substantial advances in speech recognition and natural language processing are necessary, and substantial increases in processing power and battery capacity are required before this promise can be realized.

Mobile devices are unlikely to provide printed output, but a mobile sales representative will likely carry a portable printer with which to create documents for the customer. Alternatively, such documents might be transferred directly between the mobile sales representative's device and the customer's device, using a cellular, infrared, bluetooth, or other wireless connection, without a physical paper record.

Storage capacity is not really a limitation for mobile commerce; hard disk capacities of many Gbytes are available for mobile devices. Similarly, the bandwidth of cellular communication links is sufficient for commercial interactions; however, the cost of transmitting detailed graphics over a cellular link is relatively high. Therefore, a mobile commercial sales representative will likely carry, on hard disk or CD, presentations and catalogs that contain detailed graphics or video, so that they do not need to be downloaded over an expensive wireless connection.

Typical mobile devices operate with low bandwidth, too low to allow effective display of video or Web pages. Remarkable efforts have been made with i-mode services (Kinoshita, 2002; Lane, 2002) to achieve effective mobile commerce, despite bandwidth limitations. The 3G networks currently being deployed provide sufficient bandwidth for display of video and Web pages. However, the high cost of cellular communication remains a significant limitation on activities that require large amounts of information to be transmitted. Mobile commercial activities need to operate with minimal or intermittent connections and with activities conducted while disconnected.

Currently, battery power and life are also significant limitations on mobile multimedia devices, restricting the availability of processing, display, and communication. However, small, light, mobile, alcohol-based fuel cells are in prototype and demonstration. When substantial demand develops for more powerful mobile multimedia devices, more powerful batteries will become available.

NEEDS OF USERS

It is important to distinguish between the needs of sellers and buyers and, in particular, the needs of:

- Professional mobile sellers;
- Professional mobile buyers;
- Convenience purchasers.

The popular concept of mobile commerce focuses on the buyer, but buyers are motivated by convenience, and attractive, effective capabilities are required to achieve significant adoption by buyers. In contrast, sellers are motivated by need, and they are more likely to be early adopters of novel technology.

Needs of Mobile Sellers

Professional mobile sellers include insurance agents, contractors, and other sales people who make presentations on the customers' premises. In the Internet era, with customers who do not need to visit a seller to make a purchase, sellers no longer need to wait for customers but need to become mobile to find customers wherever they can be found. Mobile sellers require support for contact information, appointments, scheduling, and reminders. PC-based tools provide such services, although their human interfaces are not appropriate for mobile devices. Mobile sales people might also use Customer Relationship Management

(CRM) software that likely will run on a central server and will be accessed remotely by a seller using a cellular Internet connection.

The most demanding aspect of the work of a mobile sales person is the presentation to the customer. A mobile sales person lacks the large physical stock and demonstration models available at a fixed site but, instead, must depend on a computer-generated display of the product. An impressive multimedia presentation is essential for selling in an environment that is saturated with television, video, animated Web pages, and the like. Thus, a mobile sales person can be expected to carry a display device (a laptop computer or a projection display), with presentations and catalogs stored on hard disk or CD. Significant effort is required to make a computer-hosted catalog as convenient to use as a conventional paper catalog, but a large computer-hosted catalog is more convenient to carry, can be searched, can contain animations, and can be updated more easily and more frequently.

Access to a catalog or other presentation material hosted on a central server is unattractive because of the cost and time of downloading detailed graphic presentations over an expensive wireless link. However, a mobile sales person needs a cellular Internet communication with the central server to query inventory, pricing, and delivery; to enter sales orders; to make reservations; and to schedule fulfillment of the sale. The mobile sales person also needs to generate proposals and contracts on the mobile device and print them for the customer. Many customers will accept electronic delivery of proposals and exchange of contracts; however, some customers will require paper copies. and, thus, the mobile sales person must carry a printer.

In summary, a laptop computer with a cellular modem and a portable printer, possibly augmented by a projection display for multimedia presentations, can satisfy the needs of a mobile sales person.

Needs of Mobile Buyers

The direct mobile buyer analog of the mobile sales person, a buyer who visits sellers to purchase goods, such as a buyer who visits ranchers to purchase livestock or visits artists to purchase paintings, is unlikely to develop. Such sellers have already discovered the use of the Internet to sell their products at higher prices than such a visiting buyer would offer.

Professionals who need to purchase while mobile include contractors and travelers. When using a mobile device such as a cell phone or PDA, they are likely to limit their activities to designation of items and quantities, delivery address and date, and payment information. They are unlikely to use such mobile devices to browse catalogs and select appropriate merchandise, because of the inadequate display and input capabilities of the devices and because of the cost of cellular Internet connections. It is essential to analyze carefully the model of transactions in a specific field of commerce, and the software and interactions needed to support that model (Keng, 2002).

Current cell phones and PDAs are barely adequate in their input and output capabilities for purchase of items in the field (Buranatrived, 2002; Lee, 2003). The small display size of portable mobile devices is unlikely to change soon but can be compensated to some extent by Web pages that are designed specifically for those devices. Such mobile-friendly Web pages must be designed not only to remove bandwidth-hogging multimedia and graphics and to reduce the amount of information presented, but also to accommodate a professional who needs to order items with minimum interaction.

Web pages designed originally for high-resolution desktop computers can be downgraded automatically, so that they require less transmission bandwidth. However, such automated downgrading does not address the abbreviated interaction sequences needed by a professional using a mobile device. Most professionals would prefer to make a conventional phone call to purchase goods, rather than to use an existing mobile device.

Mobile devices such as cell phones and PDAs have inadequate input capabilities for such mobile buyers, particularly when they are used in restricted settings such as a building site or a moving truck. This problem will be alleviated by natural language voice input when it becomes good enough. Until then, professional mobile buyers might prefer to select a small set of items from the catalog, download them in advance to the mobile device using a fixed infrastructure communication link, and use a retrieval and order program specifically designed for accessing the downloaded catalog items on the mobile device.

Needs of Convenience Purchasers

Convenience purchasers expect simpler human interfaces and lower costs than professional sellers or buyers (Tarasewich, 2003). For the convenience purchaser, because of the poor human interfaces of current mobile devices, a purely digital mobile commercial transaction is substantially less convenient and satisfying than visiting a store, making a conventional telephone call, or using the better display, easier interfaces, and lower costs of a PC to purchase over the Internet.

Convenience purchasers are most likely to purchase products that are simple and highly standardized, or that are needed while mobile. Nonetheless, mobile devices can facilitate commercial transactions in ways other than direct purchase. For example, a mobile device associated with its human owner can be used to authorize payments in a way that is more convenient than a credit card (Ogawara, 2002).

The mobile device is usually thought of as facilitating commercial transactions through mobility in space, but locating a customer in space is also an important capability (Bharat, 2003). However, location-aware services typically benefit the seller rather than the mobile purchaser, and somewhat resemble spam. A mobile device also can be used to facilitate the collection of information through time, particularly if the device is continuously present with and available to its user.

ENABLING TECHNOLOGY FOR MOBILE MULTIMEDIA

The Wireless Application Protocol, Wireless Markup Language, and Wireless Security Transport Layer discussed next are used in commercial mobile devices and enable the use of mobile multimedia for commerce.

Wireless Application Protocol

The Wireless Application Protocol (WAP) is a complex family of protocols (WAP Forum, 2004), for mobile cell phones, pagers, and other wireless terminals. WAP provides:

- Content adaptation, using the Wireless Markup Language (WML) discussed later, and the WMLScript language, a scripting language similar to JavaScript that is oriented toward displaying pages on small low-resolution displays.
- Reliability for display of Web pages provided by the Wireless Datagram Protocol (WDP) and the Wireless Session Protocol (WSP) to cope with wireless connections that are rather noisy and unreliable.
- Efficiency, provided by the WDP and the WSP through data and header compression to reduce the bandwidth required by the applications.
- Integration of Web pages and applications with telephony services provided by the WSP and the Wireless Application Environment (WAE), which allows the creation of applications that can be run on any mobile device that supports WAP.

Unfortunately, WAP's low resolution and low bandwidth are traded off against convenience of use. Because screens are small and input devices are primitive, selection of a service typically requires inconvenient, confusing, and time-consuming steps down a deep menu structure. Successful applications have been restricted to:

- Highly goal-driven services aimed at providing immediate answers to specific problems, such as, "My flight was canceled; make a new airline reservation for me."
- Entertainment-focused services, such as games, music, and sports, which depend on multimedia.

As mobile devices become more capable, WAP applications will become easier to use and more successful.

Wireless Markup Language (WML)

The Wireless Markup Language (WML), which is based on XML, describes Web pages for low-bandwidth mobile devices, such as cell phones. WML provides:

- Text presentation and layout – WML includes text and image support, including a variety of format and layout commands, generally simple and austere, as befits a small screen.

- Deck/card organizational metaphor – in WML, information is organized into a collection of cards and decks.
- Intercard navigation and linking – WML includes support for managing the navigation between cards and decks with reuse of cards to minimize markup code size.
- String parameterization and state management – WML decks can be parameterized using a state model.
- Cascading style sheets – these style sheets separate style attributes for WML documents from markup code, reducing the size of the markup code that is transmitted over a cellular link and that is stored in the memory of the mobile device.

WML is designed to accommodate the constraints of mobile devices, which include the small display, narrow band network connection, and limited memory and computational resources. In particular, the binary representation of WML, as an alternative to the usual textual representation, can reduce the size of WML page descriptions.

Unfortunately, effective display of pages on low-resolution screens of widely different capabilities requires WML pages that are specifically, individually, and expensively designed for each different mobile device, of which there are many. In contrast, HTML allows a single definition for a Web page, even though that page is to be viewed using many kinds of browsers and displays.

Wireless Transport Security Layer

Security is a major consideration in the design of systems that provide mobile multimedia for commerce. The Wireless Transport Security Layer (WTSL) aims to provide authentication, authorization, confidentiality, integrity, and non-repudiation (Kwok-Yan, 2003; WAPForum, 2004; Wen, 2002). Major concerns are:

- Disclosure of confidential information by interception of wireless traffic, which is addressed by strong encryption.
- Disclosure of confidential information, including location information within the wireless service provider's WAP gateway, which can be handled by providing one's own gateway, although most users might prefer to rely on the integrity of the wireless service provider.
- Generation of transactions that purport to have been originated by a different user, which can be handled by Wireless Identity Modules (WIMs). A WIM, which is similar to a smart card and can be inserted into a WAP-enabled phone, uses encryption with ultra-long keys to provide secure authentication between a client and a server and digital signatures for individual transactions. WIMs also provide protection against interception and replay of passwords.
- Theft and misuse of the mobile device, or covert Trojan horse code that can extract encryption keys, passwords, and other confidential information from the mobile device, which is handled by WIMs that can be but probably will not be removed from the mobile device for safe keeping, and that can themselves be lost or stolen.

WTSL is probably provides adequate security for most commercial mobile multimedia transactions, and is certainly more secure than the vulnerable credit card system that is used today for many commercial transactions.

CONCLUSION

Mobile multimedia will be a significant enabler of commerce in the future, as mobile devices become more capable, as multimedia provides more friendly user interfaces and experiences for the users, and as novel business models are developed. Great care must be taken to design services for mobile multimedia commerce for the benefit of the mobile user rather than the sellers of the service. Natural language voice input and intelligent software agents will increase the convenience of use and, thus, the popularity of mobile devices for commercial transactions.

It is not easy to predict innovations in commercial transactions; the most revolutionary and successful innovations are the most difficult to predict, because they deviate from current practice. In particular, mobile multimedia devices can be expected to have major, but unforeseeable, effects on social interactions between people, as individuals and in groups. Novel forms of social interaction will inevitably engender new forms of commercial transactions.

REFERENCES

Bharat, R., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61-65.

Buranatrived, J., & Vickers, P. (2002). An investigation of the impact of mobile phone and PDA interfaces on the usability of mobile-commerce applications. *Proceedings of the IEEE 5th International Workshop on Networked Appliances*, Liverpool, UK.

Chung-wei, L., Wen-Chen, H., & Jyh-haw, Y. (2003). A system model for mobile commerce. *Proceedings of the IEEE 23rd International Conference on Distributed Computing Systems Workshops*, Providence, Rhode Island.

Eunseok, L., & Jionghua, J. (2003). A next generation intelligent mobile commerce system. *Proceedings of the ACIS 1st International Conference on Software Engineering Research and Applications*, San Francisco, California.

Hanebeck, H.C.L., & Raisinghani, M.S. (2002). Mobile commerce: Transforming vision into reality. *Journal of Internet Commerce, 1*(3), 49-64.

Jarvenpaa, S.L., Lang, K.R., Takeda, Y., & Tuunainen, V.K. (2003). Mobile commerce at crossroads. *Communications of the ACM, 46*(12), 41-44.

Keng, S., & Zixing, S. (2002). Mobile commerce applications, in supply chain management. *Journal of Internet Commerce, 1*(3), 3-14.

Kinoshita, M. (2002). DoCoMo's vision on mobile commerce. *Proceedings of the 2002 Symposium on Applications and the Internet*, Nara, Japan.

Kwok-Yan, L., Siu-Leung, C., Ming, G., & Jia-Guang, S. (2003). Lightweight security for mobile commerce transactions. *Computer Communications, 26*(18), 2052-2060.

Lane, M.S., Zou, Y., & Matsuda, T. (2002). NTT DoCoMo: A successful mobile commerce portal. *Proceedings of the 7th International Conference on Manufacturing and Management*, Bangkok, Thailand.

Lee, Y.E., & Benbasat, I. (2003). Interface design for mobile commerce. *Communications of the ACM, 46*(12), 48-52.

Ogawara S., Chen, J.C.H., & Chong P.P. (2002). Mobile commerce: The future vehicle of e-payment in Japan? *Journal of Internet Commerce, 1*(3), 29-41.

Ortiz, G.F., Branco, A.S.C., Sancho, P.R., & Castillo, J.L. (2002). ESTIA—Efficient electronic services for tourists in action. *Proceedings of the 3rd International Workshop for Technologies in E-Services*, Hong Kong, China.

Senn, J.A. (2000). The emergence of m-commerce. *IEEE Computer, 33*(12), 148-150.

Tarasewich, P. (2003). Designing mobile commerce applications. *Communications of the ACM, 46*(12), 57-60.

Urbaczewski, A., Valacich, J.S., & Jessup, L.M. (2003). Mobile commerce opportunities and challenges. *Communications of the ACM, 46*(12), 30-32.

Varshney, U., Vetter, R.J., & Kalakota, R. Mobile commerce: A new frontier. *IEEE Computer, 33*(10), 32-38.

WAP Forum. (2004). *http://www.wapforum.com*

Wen, H.J., & Gyires, T. (2002). The impact of wireless application protocol (WAP) on m-commerce security. *Journal of Internet Commerce, 1*(3), 15-27.

KEY TERMS

Cellular Communication: Wireless communication between mobile devices (e.g., cell phones, PDAs, and PCs) and fixed base stations. The base stations serve relatively small areas of a few square miles (called cells) and are interconnected by fixed telecommunication infrastructure that provides connection with other telecommunication systems. As a mobile device passes from one cell to another, one base station hands off the communication with the device to another without disrupting communication.

Mobile Commerce: Commercial transactions in which at least one party of the transaction uses a mobile wireless device, typically a cell phone, a

PDA, or a PC equipped with a cellular modem. A PC can conduct a commercial Internet transaction using a WiFi connection to a base station, but because WiFi connections currently provide limited mobility, for this article, WiFi transactions are regarded as standard Internet transactions rather than mobile commerce.

Mobile Devices: Computing and communication devices, such as cell phones, PDAs, and PCs equipped with cellular modems. Mobile devices are lightweight and wireless so users can carry them and move about freely.

Mobile Multimedia: Use of audio and/or video in addition to text and image pages. The low bandwidth and high cost of mobile cellular connections discourage the use of video. Spoken natural language input and output is a promising but difficult approach for improving the ease of use of mobile devices for commercial transactions.

Wireless Application Protocol (WAP): The Wireless Application Protocol is an application-level communication protocol that is used to access services and information by hand-held devices with low-resolution displays and low bandwidth connections, such as mobile cell phones.

Wireless Markup Language (WML): A Web page description language derived from XML and HTML, but specifically designed to support the display of pages on low-resolution devices over low-bandwidth connections.

Wireless Transport Security Layer (WTSL): A high-security, low-overhead layer that operates above WDP and below WSP to provide authentication, authorization, confidentiality, integrity, and non-repudiation.

Mobile Radio Technologies

Christian Kaspar
Georg-August-University of Goettingen, Germany

Svenja Hagenhoff
Georg-August-University of Goettingen, Germany

INTRODUCTION

Mobile radio technologies have been subject to speculations in recent years. The initial euphoria about opportunities and market potentials of mobile services and applications has mainly been caused by growth expectations in the field of non-voice-orientated services. For the year 2003 optimistic analysis of the market development already predicted an expected total volume for the European sales of more than 23 billion Euros (Müller-Verse, 1999). But such expectations appear to be hardly achievable. In 2003 the German Ministry of Labour and Economics merely show a sales volume of US$71 million for Europe. Based on this number, an increase up to US$119 million until 2007 is predicted (Graumann/Köhne, 2003).

Due to the lack of successful business and product concepts, the gap between expectation and reality leads to insecurity about the opportunities of mobile commerce. This insecurity is mainly caused by the continuing high complexity and dynamic of mobile technologies. Therefore, particular aspects of mobile technologies as a basis of promising business concepts within mobile commerce are illustrated in the following. In order to solve insecurity problems, the application possibilities of present mobile technologies need to be analyzed on three different levels: First on the network level, whereas available technology alternatives for the generation of digital radio networks need to be considered; secondly, on the service level in order to compare different transfer standards for the development of mobile information services; thirdly, on the business level, in order to identify valuable application scenarios from the customer point of view. The following analysis considers alternative technologies on the network and service level in order to determine application scenarios of mobile technologies in the last chapter.

DIGITAL RADIO NETWORKS

In the past the analysis of mobile radio technology has often been limited to established technology standards as well as their development in the context of wide-area communication networks. Today it is recognizable that wireless technologies that have been developed for networks within locally limited infrastructures represent good and cheap alternatives to wide-area networks (Webb, 2001). Thus, in the following three alternatives, architecture and technology are represented.

General Basics of Mobile Radio Technology

Generally, connections within mobile radio networks can be established between mobile and immobile stations (infrastructure networks) or between mobile stations (ad-hoc networks) only (Müller, Eymann, & Kreutzer, 2003). Within the mobile radio network, the immobile transfer line is displaced by an unadjusted radio channel. In contrast to analogous radio networks, where the communication signal is directly transferred as a continuing signal wave, within the digital radio network the initial signal is coded in series of bits and bytes by the end terminal and decoded by the receiver.

The economically usable frequency spectrum is limited by the way of usage as well as by the actual stage of technology and therefore represents a shortage for mobile radio transmissions. Via so called "multiplexing", a medium can be provided to different users by the division of access area, time, frequency, or code (Müller, Eymann, & Kreutzer, 2003; Schiller, 2003).

In contrast to fixed-wire networks within radio networks, the signal spread takes place directly similar to light waves. Objects within the transfer area can

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

interfere with the signal spread that is why there is the danger of a signal deletion within wireless transmission processes. In order to reduce such signal faults, spread spectrum techniques distribute the initial transmission bandwidth of a signal onto a higher bandwidth (Schiller, 2003). The resulting limitation of available frequency can be minimized by the combination of spread spectrum techniques with multiple access techniques. Those forms of combination are represented, for example by the *Frequency Hopping Spread Spectrum* (FHSS), where each transmitter changes the transfer frequency according to a given hopping sequence, or the *Direct Sequence Spread Spectrum* (DSSS), where the initial signal spread is coded by a predetermined pseudo random number.

Wireless Local Area Networks (IEEE 802.11)

The developers of the 802.11 standards aimed at establishing application and protocol transparency, seamless fixed network integration and a worldwide operation ability within the license-free ISM (Industrial, Scientific and Medical) radio bands (Schiller, 2003). The initial 802.11 standard of 1997 describes three broadcast variants: one infrared variant uses light waves with wave-lengths of 850-950 nm and two radio variants within the frequency band of 2.4 GHz which are economically more important (Schiller, 2003). Within the designated spectrum of the transfer power between a minimum of 1mW and a maximum of 100mW in Europe the radio variants can achieve a channel capacity of 1-2 Mbit/s. Following the 802.3 (Ethernet) and 802.4 (Token Ring) standards for fixed-wire networks the 802.11 standard specifies two service classes (IEEE, 2001): an asynchronous service as a standard case analogous to the 802.3 standard and an optional, temporally limited synchronous service. Typically WLANs operate within the infrastructure modus whereby the whole communication of a client takes place via an access point. The access point supplies every client within its reach or serves as a radio gateway for adjoining access points.

Developments of the initial standards are mainly concentrated on the area of the transfer layer (Schiller, 2003). Within the 802.11a standard the initial 2.4 GHz band is displaced by the 5 GHz band, allowing a capacity of up to 54 Mbit/s. In contrast to this, the presently most popular standard 802.11b uses the encoded spread spectrum technique DSSS. It achieves a capacity up to 11 Mbit/s operating within the 2.4 GHz band.

Wireless Personal Area Networks (Bluetooth)

In 1998 Ericsson, Nokia, IBM, Toshiba, and Intel founded a "Special Interest Group" (SIG) for radio networks for the close-up range named "Bluetooth" (SIG, 2004). Like WLAN networks, Bluetooth devices transfer within the 2.4 GHz ISM bands, which is why interferences may occur between both network technologies. In general, 79 channels are available within Bluetooth networks. FHSS is implemented with 100 hops per second as spread spectrum technique (Bakker & McMichael, 2002). Devices with identical hop sequences constitute a so-called "pico-network". Within this network, two service categories are specified: a synchronous, circuit-switched method and an asynchronous method. Within the maximum transfer power of 10mW Bluetooth devices can reach a transfer radius of 10m up to a maximum of 100m and a data capacity of up to 723Kbit/s (Müller, Eymann, & Kreutzer, 2003).

The main application areas of Bluetooth technologies are the connection of peripheral devices like computer mouse, headphones, automotive electronics, and kitchen equipment or the gateway function between different network types like the cross linking of fixed-wire networks and mobile radio devices (Diederich, Lerner, Lindemann, & Vehlen, 2001). Generally, Bluetooth networks are therefore linked together as ad-hoc networks. Ad-hoc networks do not require decided access points; mobile devices communicate equally and directly with devices within reach. Among a network of a total maximum of eight terminals, exactly one terminal serves as a master station for the specification and synchronization of the hop frequency (Haartsen, 2000; Nokia, 2003). Bluetooth devices can be involved in different piconetworks at the same time but are not able to communicate actively within more than one of these networks at a particular point in time. These overlapping network structures are called scatter-networks.

Network Standards for Wide-Area Communication Networks

In 1982, the European conference of post and communication administration founded a consortium for the coordination and standardization of a future pan-European telephone network called "Global System for Mobile Communications" (GSM Association, 2003; Schiller, 2003). At the present, there are three GSM based mobile networks in the world with 900, 1800, and 1900 MHz, which connect about 800 million participants in 190 countries at the moment (GSM Association, 2003). In Europe, the media access of mobile terminals onto the radio network takes place via time and frequency multiplex on an air interface. This interface obtains 124 transmission channels with 200 kHz each within a frequency band of 890 to 915 MHz (uplink) or 935-960 MHz (downlink; Schiller, 2003). Three service categories are intended:

- Carrier services for data transfer between network access points; thereby circuit-switched as well as package-switched services with 2400, 4800 and 9600 Bit/s synchronous or 300-1200 Bit/s asynchronous are specified.
- Teleservices for the voice communication with initially 3.1 KHz and for additional non-voice applications like fax, voice memory, and short message services;
- Additional services like telephone number forwarding, rejection, knocking, and so on.

The architecture of an area-wide GSM network is more complex compared to local radio variants and consists of three subsystems (Müller, Eymann, Kreutzer, 2003; Schiller, 2003):

1. The radio subsystem (RSS) is an area wide cellular network, consisting of different base station subsystems (BSS). A BSS obtains at least one base station controller (BSC) which controls different base transceiver stations (BTS). Generally a BTS supplies one radio cell with a cell radius of 100m up to a maximum of 3 km.
2. The network subsystem (NSS) builds the main part of the GSM network and obtains every administration task. Their core element is the mobile switching center (MSC) which assigns a signal within the network to an authenticated participant. The authentication takes place based on two databases. Within the home location register (HLR), any contract-specific data of a user as well as his location are saved; within the visitor location register (VLR) which is generally assigned to a MSC, every participant who is situated within the actual field of responsibility of the MSC is saved.
3. The control and monitoring of networks and radio subsystems takes place via an operation and maintenance system (OMC). The OMC is responsible for the registration of mobile stations and user authorizations and generates participant specific authorization parameters.

The main disadvantage of GSM networks is the low channel capacity within the signal transfer. A lot of developments aim at the reduction of this limitation (Schiller, 2003): Within the high-speed circuit-switched data (HSCSD) method, different time slots are combined for one circuit-switched signal. The general packet radio service (GPRS), is a package-switched method that combines different time slots like the HSCSD method, but it occupies channel capacities only if the data transfer takes place. GPRS requires additional system components for network subsystems and theoretically allows a transfer capacity of 171.2 kBit/s.

The universal mobile telecommunication service (UMTS) represents an evolutionary development of GSM. The development aims at a higher transfer capacity for data services with a minimum data rate of up to 2 Mbit/s for metropolitan areas. The core element of the development is the enhanced air interface called universal terrestrial radio access (UTRA). This interface uses a carrier frequency with a bandwidth of about 1.9 to 2.1 GHz and uses a broadband CDMA technology with the spread spectrum technique DSSS.

TECHNOLOGIES FOR MOBILE INFORMATION SERVICES

The network technologies introduced above just represent carrier layers and do not enable an exchange of data on the service level on their own. Therefore some data exchange protocol standards for the development of mobile services are intro-

duced in the following. Two conceptually different methods are distinguished: the WAP model and the Bluetooth model.

WAP

Though the exchange of data within mobile networks can generally take place based on HTTP, TCP/IP, and HTML, especially the implementation of TCP within mobile networks can cause problems and may therefore lead to unwanted drops of performance (Lehner, 2003). Bearing this in mind in 1997 a cellphone manufacturer consortium developed the wireless application protocol (WAP) which aims at improving the transfer of Internet contents and data services for mobile devices. WAP represents a defacto standard which is monitored by a panel, the so-called WAP Forum, introduced by Ericsson, Motorola, and Nokia.

WAP acts as a communication platform between mobile devices and a WAP gateway. The gateway is a particular server resembling a proxy server that translates WAP enquiries into HTTP messages and forwards them to an internet content server (Deitel, Deitel, Nieto, & Steinbuhler, 2002).

In fact, WAP includes a range of protocols that support different tasks for the data transfer from or to mobile devices containing protocols for the data transfer between WAP gateway and user equipment as well as the markup language WML (Lehner, 2003). Figure 2 shows the layers of WAP compared to the ISO/OSI and the TCP/IP model.

Bluetooth

The developers of Bluetooth aimed at guaranteeing a cheap, all-purpose connection between portable devices with communication or computing capabilities (Haartsen, Allen, Inouye, Joeressen, & Naghshineh, 1998). In contrast to WLAN or UMTS the Bluetooth specification defines a complete system that ranges from the physical radio layer to the application layer. The specification consists of two layers: the technical core specification that describes the protocol stack and the application layer with authorized profiles for predefined use cases.

Within the architecture of the Bluetooth protocol stack two components are distinguished (Figure 3): the Bluetooth host and the Bluetooth controller. The Bluetooth host is a software component as a part of the operating system of the mobile device. The host is usually provided with five protocols which enable an integration of Bluetooth connections with other specifications. The Logical Link Control and Adoption Protocol (L2CAP) enable a multiple access of different logical connections of upper layers to the radio frequency spectrum. The identification of available Bluetooth services takes place via the service discovery protocol (SDP). Existing data connections like point-to-point connections or WAP services are transferred either via RFCOMM or via the Bluetooth Encapsulating Protocol (BNEP). RFCOMM is a basic transport protocol which emulates the functionality of a serial port. BNEP gathers packages of existing data connections and sends them directly via the L2CAP. The Object Exchange Protocol (OBEX) has been adapted for Bluetooth from the infrared technology for the transmission of documents like vCards.

Bluetooth profiles represent usage models for Bluetooth technologies with specified interoperability for predefined functions. Bluetooth profiles underlie a strict qualification process executed by the SIG. General transport profiles (1-4) and application profiles (5-12) for particular usage models are distinguished (Figure 4).

1. **The Generic Access Profile** (GAP) specifies generic processes for equipment identification, link management, and security.
2. **The Service Discovery Application Profile** (SDAP) provides functions and processes for the identification of other Bluetooth devices.
3. **The Serial Port Profile** (SPP) defines the necessary requirements of Bluetooth devices for the emulation of serial cable connections based on RFCOMM.

Figure 1. WAP interaction model

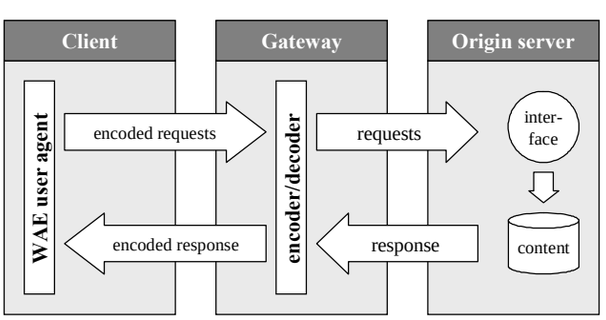


Figure 2. WAP protocol stack vs. TCP/IP

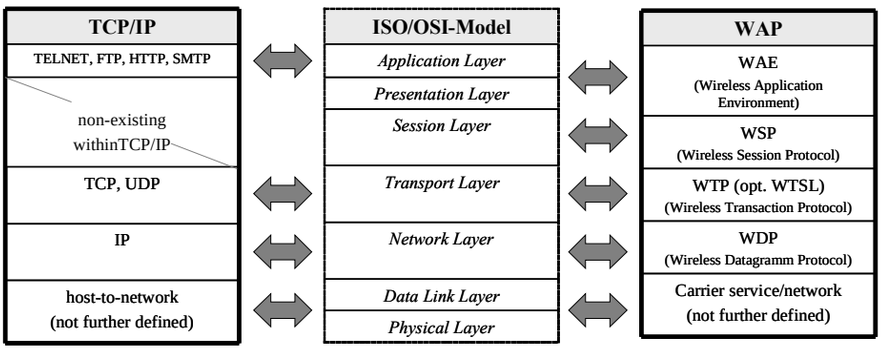


Figure 3. Bluetooth protocol stack

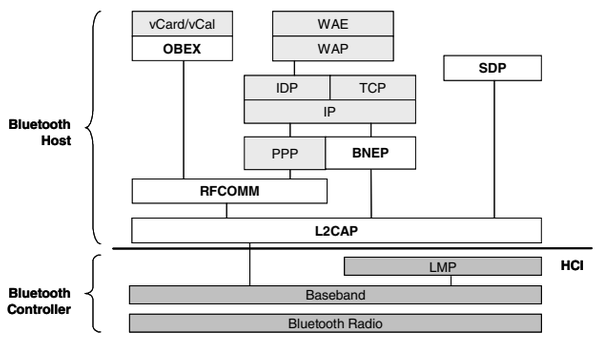
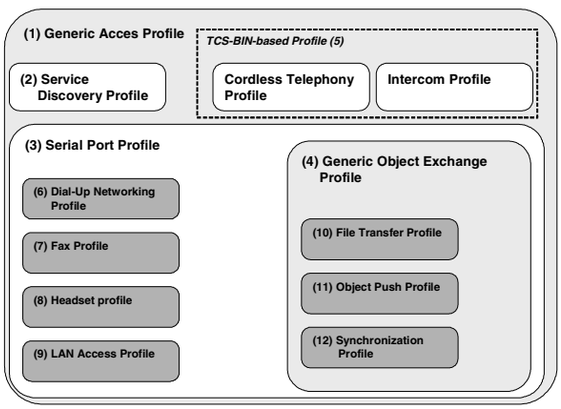


Figure 4. Bluetooth profiles



4. The **Generic Object Exchange Profile** defines processes for services with exchange functions like synchronization, data transfer or push services.

On the one hand, usage model-oriented profiles include typical scenarios for cable alternatives for the communication between devices within the short distance area. Examples are the utilization of the mobile phone as cordless telephone or intercom (5), as a modem (6), as fax (7), the connection to a headset (8) or the network connection within a LAN (9). On the other hand profiles are offered for the exchange of documents (10), for push services (11) and for the synchronization for example to applications within computer terminals (12).

CONCLUSION

The main assessment problem in the context of commercial marketing of the mobile-radio technologies is derived from the question if these technologies can be classified as additional evolutionary distribution technologies within electronic commerce only or if it is a revolutionary branch of the economy (Zobel, 2001). Against this problematic background, two general utilization scenarios need to be distinguished:

(1) In addition to conventional Ethernet, the WLAN technology enables a portable access to data networks based on TCP/IP via an air interface. The regional flexibility of the data access is the only advantage. No more additional benefits like push services or functional equipment connections and therefore no fundamental new service forms can be realized.

(2) The standards of the generation 2.5 (like GSM/GRPS) within the license bound mobile networks already enable a reliable voice communication and obtain at least sufficient capacities for the data traffic. Problems are caused by generally high connection expenses which is why the mobile network technology is no real alternative for fixed networks in the medium term. Instead of a portable data network access the scenario of particular information services

for the requirements of mobile users that generate an additional added value in contrast to stationary network utilization seems more plausible. The localization of a mobile device which for example is possible within GSM via the residence register (HLR/VLR) or the distinct identification via SIM card has been mentioned as added value factors. With regard to the possibility of push services which are for example positioned within the WAP specification, services are possible that identify the actual position of a device and automatically provide a context adapted offer for the user.

The Bluetooth technology has initially been generated as a cable alternative for the connection of terminals close by and can therefore hardly be classified as one of the two utilization scenarios. Due to the small size and little power consumption of Bluetooth systems, their implementation is plausible within everyday situations as ubiquitous cross linking technology. Thereby, two utilization scenarios are imaginable. On the one hand, the cross linking between mobile user equipment as an alternative for existing infrastructure networks for data as well as for voice communication is conceivable; on the other hand easy and fast connections between user equipment and stationary systems like within the context of environment information or point-of-sale terminals are imaginable.

Thus, it is obvious that particular fields of application seem plausible for each mobile network technology. It is therefore obvious that data supported mobile services obtain a significant market potential. However a repetition of the revolutionary change like it has been caused by the internet technologies seems rather unlikely for the mobile network technologies. Mobile network technologies can provide an additional value in contrast to conventional fixed networks and new forms of service can be generated using existing added value factors of the mobile network technology. However, these advantages are mainly efficiency based, for example the enhanced integration ability of distributed equipments and systems or a more comfortable data access.

REFERENCES

Bakker, D. & McMichael Gilster, D. (2002). *Bluetooth end to end*. John Wiley and Sons.

Deitel, H., Deitel, P., Nieto, T., & Steinbuhler, K. (2002). *Wireless Internet & mobile business – How to program*. New Jersey: Prentice Hall.

Diederich, B., Lerner, T., Lindemann, R., & Vehlen, R. (2001). Mobile Business. Märkte, Techniken, Geschäftsmodelle. Gabler, Wiesbaden 2001.

Graumann, S. & Köhne, B. (2003). Monitoring Informationswirtschaft. 6 Faktenbericht 2003, im Auftrag des Bundesministeriums für Wirtschaft und Arbeit. *http://www.bmwi.de/Redaktion/Inhalte/Downloads/6-faktenbericht-vollversion,templateId=download.pdf* [2004-05-31].

GSM Association (2003). Membership and Market Statistics, March 2003; *http://www.gsmworld.com/news/statistics/feb03_stats.pdf [2004-05-31]*

Haartsen, J., Allen, W., Inouye, J., Joeressen, O., & Naghshineh, M. (1998). Bluetooth: Vision, goals, and architecture. *Mobile Computing and Communications Review, 2*(4), 38-45.

IEEE (2001). Functional Requirements - IEEE Project 802. *http://grouper.ieee.org/groups/802/802_archive/fureq6-8.html [2004-05-31]*

Jaap, C.H. (2000). The Bluetooth Radio System. *IEEE Personal Communications, 7*, 28-36.

Kumar, B., Kline, P., & Thompson T. (2004). *Bluetooth application programming with the JAVA APIs. The Morgan Kaufmann series in networking*. San Francisco: Elsevier.

Lehner, F. (2003). *Mobile und drahtlose Informationssysteme*. Berlin: Springer Verlag.

Müller, G., Eymann, T., & Kreutzer, M. (2003). Telematik- und Kommunikationssysteme in der vernetzten Wirtschaft. Lehrbücher Wirtschafts informatik. Oldenbourg Verlag. München 2003.

Müller-Verse, F. (1999). Mobile Commerce Report. Durlacher Research Report. *http://www.durlacher.com/downloads/mcomreport.pdf [2004-05-31]*

Nokia (2003). Bluetooth Technology Overview. Version 1.0. April 4, 2003. *http://ncsp.forum.nokia.com/downloads/nokia/documents/Bluetooth_Technology_Overview_v1_0.pdf [2004-05-31]*

Schiller, J. (2003). *Mobile communications*. Upper Saddle River, NJ: Addison-Wesley.

SIG (2004). The Official Bluetooth Membership Site. *https://www.bluetooth.org [2004-05-31]*

Webb, W. (2001). *The future of wireless communications*. Boston: Artech House.

Zobel, J. (2001). Mobile business and m-commerce. München : Carl Hanser Verlag.

KEY TERMS

Bluetooth: A specification for personal radio networks, named after the nickname of the Danish king Harald who united Norway and Denmark in the 10th century.

Circuit Switching: Circuit-switched networks establish a permanent physical connection between communicating devices. For the time of the communication this connection can be used exclusively by the communicating devices.

GSM: In 1982 The European conference of post and communication administration founded a consortium for the coordination and standardization of a future pan-European telephone network called "Group Spécial Mobile" that was renamed as "Global System for Mobile Communications" later.

Local Area Radio Network: Mobile radio networks can either be built up as wide area networks consisting of several radio cells or as local area networks usually consisting of just one radio cell. Depending on the signal reach of the used transmission technology a local area network can range from several meters up to several hundred meters.

Multiplex: Within digital mobile radio networks, three different multiplexing techniques can be applied: *Time Division Multiple Access* (TDMA), *Frequency Division Multiple Access* (FDMA) and *Code Division Multiple Access* (CDMA).

Packet Switching: Packet-switched networks divide transmissions into packets before they are sent. Each packet can be transmitted individually and is sent by network routers following different routes to its destination. Once all the packets forming the initial message arrive at the destination they are recompiled.

Personal Area Radio Network: Small sized local area network can also be named as wireless personal area network or wireless close-range networks.

Wide Area Radio Network: A wide area radio network consists of several radio transmitters with overlapping transmission ranges.

Mobility over Heterogeneous Wireless Networks

Lek Heng Ngoh
*Institute for Infocomm Research, A*STAR, Singapore*

Jaya Shankar P.
*Institute for Infocomm Research, A*STAR, Singapore*

INTRODUCTION

Accessing wireless services and application on the move has become a norm among casual or business users these days. Due to societal needs, technological innovation, and networks operators' business strategies, there has been a rapid proliferation of many different wireless technologies. In many parts of the world, we are witnessing a wireless ecosystem consisting of wide-area, low-to-medium-bandwidth network based on access technologies such as GSM, GPRS, and WCDMA, overlaid by faster local area networks such as IEEE 802.11-based Wireless LANs and Bluetooth pico-networks. One notable advantage of wide-area networks such as GPRS and 3G networks is their ability to provide access in a larger service area. However, a wide-area network has limited bandwidth and higher latency. 3G systems promise a speed of up to 2Mbps per cell for a non-roaming user. On the other hand, alternative wireless technologies like WLAN 802.11and Personal area network (PAN) using Bluetooth technology have limited range but can provide much higher bandwidth. Thus, technologies like WWAN and WLAN provide complementary features with respect to operating range and available bandwidth. Consequently, the natural trend will be toward utilizing high bandwidth data networks such as WLAN, whenever they are available, and to switch to an overlay service such as GPRS or 3G networks with low bandwidth, when coverage of WLAN is not available. Adding to the existing public networks, some private institutions (i.e., universities) have joined the fray to adopt wireless infrastructure to support mobility within their premises, thus adding to the plethora of wireless networks. With such pervasiveness, solutions are required to guarantee end-user terminal mobility and maintain always-on session connections to the Internet. To achieve this objective, an end device with several radio interfaces and intelligent software that would enable the automatic selection of networks and resources is necessary (Einsiedler, 2001; Moby Dick, 2003).

Related Technical Challenges

While this article focuses on how an IP-based mobile node can remain connected to the Internet as it moves across different network technologies, for practical and commercial Internet deployment, functions such as access authentication, security, and metering (for charging purposes) also need to be integrated with these mobility functions. In addition, in order to support the needs of cost-savvy users and future real-time applications such as VOIP and video conferencing, functions such as intelligent interface/network selection, fast and seamless hand-over, context transfer, QoS provisioning, differentiation, and others yet to be thought of need to be integrated. Moreover, specific variants of each of these functions, tailor-fitted to specific access technologies, may have to co-exist on mobile stations equipped with multiple access technologies.

In the remainder of this article, the various technical challenges are elaborated upon through a number of commercially available and research solutions presented in detail.

POSSIBLE SOLUTIONS

In general, a seamless mobility solution can be achieved by using two main approaches. The first is based on Mobile IP Technology. The second approach is a

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Central Server Based solution. In the Mobile IP-based solution, there are specific solutions for IPv4 and IPv6 (Deering, 1998), respectively. To provide a comprehensive review of these solutions, the rest of this article is organized as follows: first, the mobile IPv6 solution is explained (Johnson, 2003); this is followed by two IPv4 solutions, namely the mobile IPv4 and central server based solution.

Mobile IPv6-Based Solutions

In the current state of the art, Mobile IPv6 (Johnson, 2003; Koodli, 2003) makes it possible for an IPv6 mobile node (MN) to remain connected to the Internet as it changes its network point of attachment. However, from a network provider's point of view, in addition to the mobility function, the system needs to be integrated with additional functions that allows them to authenticate, provision/select, and maintain suitable network resources, charge the MN for usage of their infrastructure, and so forth. From a network user's point of view, the MN needs to be smart enough to automatically select and hand off to networks that best suit its policies as and when they become available. Additionally, in the case of multi-homed MNs, it should be smart enough to automatically route traffic through the interface that best suits its policies (Kenward, 2002; Loughney, 2003; Thomson, 1998). All these need to be done in a seamless manner by the MN, where possible.

An example of a IPv6-based mobility solution is the AMASE (Advance Mobile Application Support Environment) project (Jayabal, 2004). It is aimed at providing a middleware for mobile devices that will allow users to move from one network to another and still have access to rich multimedia services in a seamless manner. One of the key features of this middleware is the intelligent abstraction of the underlying networks and network resources that are handled by a module called UAL (Universal Adaptation Layer). This entity is the client part of a mobility and resource management framework, which provides the mobility function in AMASE while, at the same time, facilitating the other additional functions to be carried out, as mentioned previously. The components of AMASE are elaborated in the following section:

The Mobile Node (MN)

The MN is a physical entity installed with the following AMASE logical components:

- **Universal Aadaptation Layer (UAL):** The UAL consists of the Mobility Management (MM) framework and a simple user –policy-based local network resource and handover management function (SLRM). It is responsible for the automatic link/network discovery and for IPv6 roaming mobility of the MN. The MM framework is designed so that it can be extended to facilitate other additional functions such as the URP, while managing the IPv6 mobility of the MN. The DHC6C module provides programmatic interfaces to the MMC for triggering DHCPv6 procedures, receiving DHCPv6 events, and sending and receiving MM and MM function-specific messages. The LM consists of network device-specific components that abstract the control, status reporting, and parameters of available links in each of the network devices governed by the UAL to present a uniform programmatic interface to MMC or other MMC extended functions. Presently, a generic LM module for single-link interfaces (i.e., wall-plug Ethernet or GPRS) and a signal strength and hysteresis-based LM for 802.11 interfaces (LM80211) are implemented.
- **DHCPv6 Client (DHC6C):** AMASE-enabled mobile node implements the Dynamic Host Configuration Protocol for IPv6 (DHCPv6) (Droms, 2003) to achieve stateful address autoconfiguration. Apart from obtaining IPv6/v4 address(es) from the network, DHCPv6 provides a flexible mechanism for the mobile node to request configuration parameters from the server, which is the underlying signaling protocol used in AMASE to obtain several AMASE-specific configurations. The design of the DHCPv6 client allows AMASE-modules (e.g. URP, MM) or applications to react according to specific DHCPv6-specific events. The design also allows AMASE-modules to have control of the behavior of the DHCPv6 state-machine.

- **URP Client (URPC):** The URPC (User Registration Protocol Client) (Forsberg, 2003) is a software module that implements authentication of the user/machine via a AAA (Authentication, Authorization, and Accounting) framework. It also is responsible for configuring the IPSec tunnel protecting the wireless last-hop.

Smartcard Module (SC)

The AMASE smartcard module (SC) provides the mobile node a secure means to authenticate users to the network/service. The module provides applications or AMASE modules (e.g., URP) to register and handle specific smartcard events (e.g., smartcard removal/insertion) through a well-defined interface. More importantly, the interface provides a means to access the underlying services such as cryptography, encryption/decryption, and session-key generation, which are performed inside a Javacard-enabled smartcard.

Shipworm Client (SPWMC)

The shipworm (Huitema, 2001) client is an IPv6/IPv4 interworking function residing on MN. The shipworm client acts as an IPv6-enabled network device. It converts all IPv6 traffic into IPv4 traffic and sends to shipworm server through an IPv4 connection.

QoS Module

The QoS module in MN performs the following tasks: When MN registers to the network, it downloads QoS-related policies from Resource Allocation Policy Decision Point (RA-PDP) and installs them on MN. The policies include, for example, different classes of services the network can provide to MN and, optionally, their pricing information, network access preferences and restrictions, and so forth. The QoS module interacts with user application programs through a set of APIs for QoS-required network connection setup requests. It then interacts with RA-PDP to request and reserve the necessary network resources for the connection. When the application program closes the network connection, it then informs RA-PDP to release the reserved network resources. During the handoff process, the QoS module collaborates with UAL and RA-PDP to provide QoS-enabled handoff.

The Mobility Gateway (MG)

The MG is an AMASE access router installed with the following AMASE logical components:

- **Mobility Management Gateway Component:** The MM on the MG is the gateway counterpart to the MM on the MN. It consists of the mobility management core states and procedures module (MMC), a highly interfaceable server-part DHCPv6 module (DHC6S) and several configuration hooks into the IPv6 stack. The MMC is the main controller of the MM and is designed so that it can be extended to facilitate other additional functions such as URP and network-controlled, fast, and anticipative handover management while managing mobility of the MNs.
- **DHCPv6 Server (DHC6S):** The DHCP server is configured to pass configuration parameters such as IPv6/4 addresses to the MN. The server also provides stateless DHCPv6 services to MNs, which doesn't require addresses from the server. In the stateful mode, the server maintains per-MN configuration information such as addresses. Similar to the design of DHC6C, it provides a mechanism for AMASE-modules (e.g., URP, MM) or applications to react according to specific DHCPv6-specific events. The design also allows AMASE-modules to have control of the behavior of the DHCPv6 state-machine.
- **URP Server (URPS):** The URPS (User Registration Protocol Server) is the entity that carries out user/machine authentication of an MN connecting to the network. It communicates with the URPC on the MN and terminates the URP. It interfaces with the AAA Client to carry out authentication through the AAA framework. The URPS and the AAA Attendant together are rather similar in spirit to the NAS (Network Authentication Server) in a RADIUS-based architecture (Rigney, 2000). The URPS is responsible for (a) authentication of user/MN; (b) protecting the network from unauthenticated ingress traffic; and (c) protecting the communication between an authenticated MN and the network over the last-hop wireless link.

- **AAA Client (AAAC):** AAA Client works in conjunction with AAA Server in order to provide challenge-based mutual authentication for MN. With the help of the pre-established security association among the entities such as MN, MG, AAA visited-domain server (AAAL), and AAA home-domain server (AAAH), such an interaction will facilitate MN's registration in a visited domain. To implement the challenge-based mutual authentication protocol, RADIUS protocol containing vendor-specific EAP messages are used. The authentication is bidirectional, and session keys are distributed to enable secure communication between MN and MG.
- **Resource Allocation Policy Enforcement Point (RA-PEP):** The MG acts as a Policy Enforcement Point (PEP) for the management of network resources. During registration, when an MN attaches to MG, the visited domain's RA-PDP will push a set of network resource policies to MG. The policies are typically the total bandwidth limits for each class of services for that MN. MG installs those policies on its Traffic Controller (TC) module to start regulating and enforcing the network traffic to and from the MN. In AMASE QoS design, there are three classes of services; namely, Expedite Forwarding (EF), Assured Forwarding (AF), and Best Effort (BE). When MN initially attaches to MG, only BE service is allowed. After MN successfully reserves network resources from RA-PDP for subsequent EF or AF network connections, RA-PDP will push a new set of policies to MG to allow such connections.

The Home Agent (HA)

On networks served by MGs not installed or configured to run the MIPv6 HA function, a separate machine is used to provide the same function.

Access, Authentication, and Accounting Policy Decision Point (AAA-PDP)

An AAA-PDP (Access, Authentication, and Accounting) infrastructure has been leveraged to have effective roaming between different domains. Since the AAA verification systems currently are used for interdomain roaming support and for accounting services, this infrastructure also can be used for key distribution. For the authentication system, RADIUS protocol has been selected for the AAA infrastructure due to its well-established standard and widespread existing installation. Since standard RADIUS protocol will not suffice the authentication requirements of mutual authentication and key distribution, extended RADIUS protocol with Extensible Authentication Protocol (EAP) (Rigney, 2000) attribute has been adopted for the authentication system. Within the EAP attribute, various new EAP subtypes are defined to carry the authentication-related messages.

Shipworm Server (SPWMS)

This is an IPv4-IPv6 interworking function (IWF). Shipworm server is the counterpart of shipworm client. It waits for tunneled packets from shipworm clients and forwards the IPv6 packets inside according to the IPv6 routing information. It also is responsible for tunneling the IPv6 packets, which are destined to shipworm clients to related shipworm clients.

Resource Allocation Policy-Decision-Point (RA-PDP)

The RA-PDP is a generic entity responsible for resource (e.g., bandwidth) provisioning, tracking, and admission control. After a successful authentication between MN and the network, the network's RA-PDP pushes a set of network resource policies to the MG to which MN is attached. MG then will install and enforce the policies. The contents of policies are decided in the Service Level Agreement (SLA) between MN and the network. One example could be total bandwidth limitations for each class of services the network can provide to MN. Other examples could be pricing information, network access restrictions, routing enforcement, and so forth. In a network that requires resource allocation signaling, RA-PDP also accepts Resource Allocation Requests (RARs) from MNs and makes resource allocation decisions based on current network resources, MN's SLAs, and current resource utilization of MNs. Upon a successful resource allocation, RA-PDP pushes a new set of policies to MG and allows the allocated resources to be used by MN. There are two functional entities in RA-PDP. One is a PDP that implements COPS and

COPS-PR protocol. The other is a Bandwidth Broker (BB) that does the resource allocation and tracking.

Mobile IPv4-Based Solutions

Mobile IP is an IETF standard protocol (Perkins, 2002; Calhoun, 2000) designed to allow Internet nodes to achieve seamless mobility. Mobile IP support requires a minimum of both a home agent and a mobile node; a more comprehensive solution also will involve a foreign agent acting on behalf of multiple mobile nodes. Typically, the home agent sits at the user's home network and intercepts IP datagrams from a host destined for the mobile node. The datagrams then are tunneled by the home agent and forwarded to either a foreign agent or the mobile node using a temporary IP address. Finally, the mobile node unpacks the original datagram and reinserts it into the stack, resulting in a transparent operation using the original IP addresses only. Replies to the originating host either can be sent directly from the mobile node to the host or tunneled back to the home agent, which, in turn, unpacks and forwards the replies to the host. To achieve this, a number of technical issues need to be resolved, and related IETF standard protocols are needed. These issues are elaborated as follows:

- **Tunneling Protocol:** IP datagrams to the mobile node are tunneled from the home network by the home agent to the foreign agent or mobile node directly. IP in IP tunneling defined by RFC 2003 (Perkin, 1996)is the default and mandatory tunneling protocol and is supported by a mobile node. Generic Routing Encapsulation (GRE), an optional tunneling method that can be used with mobile IP, also is supported.
- **Network Address Translation (NAT) Traversal:** The problem of traversal of Mobile IP over NAT is solved by using the Mobile IP NAT support, according to RFC 3519 (Levkowetz, 2003) Mobile IP Traversal of Network Address Translation (NAT) Devices standard.
- **Reverse Tunneling:** The default operation with mobile IP is to send replies directly to a host using standard IP routing (i.e., without tunneling or passing the datagram through the home agent). The effect is a triangular routing pattern where the host sends its datagrams to the home agent, which, in turn, tunnels them to the mobile node. Finally, the mobile node sends its datagrams directly to the original host, resulting in the triangle. However, due to various security mechanisms like ingress filtering and firewalls, this mode of operation may not work because the datagrams from the mobile node are discarded. The solution is to also tunnel and forward datagrams originating from the mobile node through the home agent. This mode of operation is called Reverse Tunneling (RFC 3024) (Montenegro, 2002).
- **Security:** Registration messages exchanged between a mobile node and its home agent are always authenticated through the use of a shared secret, which is never sent over the network. More specifically, the secret is used with keyed MD5 in prefix + suffix mode to create a 128-bit message digest of the complete registration message, not only serving to verify the sender, but also to protect the message from alterations. Replay protection is realized with timestamps. The optional Reverse Tunneling feature may be utilized if firewalls are used. A positive side effect of reverse tunneling is that the whereabouts of the mobile node are hidden from the hosts with which it communicates.

Central-Server-Based Mobility Solutions

Over the years, there have been other solutions that have been developed based on non-mobile IP but using a client-server-based solution. The non-mobile IP-based solutions can be broken down further into two main approaches. The first is based on IP and the rest on non-IP. Examples of non-IP solutions are session layer or application layer mobility. Typical examples of application layer mobility are WAP-based solution, IBM's Web sphere, and SIP-mobility (IBM's Everyplace Wireless Gateway). In the non-IP solutions, a higher layer protocol such as TCP or UDP is used to implement the mobility. In some cases, a split sessions approach is used so that sessions are terminated and restarted on the server side. The main drawback of these solutions is that end-to-end security semantics can be compromised.

The other approach is an IP layer approach (e.g., NetMotion Wireless), whereby the packets are tunneled to a server using the IP address that is obtained from the server and the IP address from the local access network. The concepts involved are analogous to the IETF-defined Mobile IP protocol. It includes client software and a mobility server whose function is similar to that of the home agent. The major difference between this and the standards approach is that the mobility solution is based on a shim, or driver, that sits between the application layer and the transport layer. Because the driver sits beneath the application layer, applications are unaware of the mobility mechanism in place. Because there is no change in the IP stack, rebuilding the operating system or replacing or enhancing the IP stack of the mobile client becomes unnecessary. A mobility server acts as a proxy for the mobile device, which is assigned an IP address that results in packets destined for the mobile node being routed to the mobility server. The mobility server knows the mobile's current location and care-of address, and is able to forward the packets. The solution requires a mobility server as well as the installation of proprietary software on the client. However, this solution does not involve a foreign agent but uses a movement-detection mechanism that is based either on link-layer triggers provided by the interface card driver or Dynamic Host Configuration Protocol discover broadcasts.

Finally, another proprietary IP-mobility solution is the PacketAir's Mobility Router (PMR) (PacketAir Networks), which must be part of the access network. This solution is radio technology agnostic and can be deployed in a variety of environments, including local- and wide-area networks. As the mobile moves to different subnets that also have PMRs, the mobile's IP address continues to be anchored at the first PMR. Tunnels between the edge routers are extended during movement, and session continuity is maintained because the mobile does not experience a change in IP address. Movement detection is accomplished by proprietary mechanisms, and the company claims sub-20-millisecond handoff rates. The PMR-based solution works when the coverage area of a mobility router is large, and, hence, the subnet that the mobile is in changes infrequently. If the base stations are considered as edge routers, however, the PMR solution does not scale well, because the tunnels would need to be extended across a number of base stations. Also, the ability to detect a change in the link at such rapid speeds (less than 20 ms) is closely tied to the access technology itself. Such triggers may not be available readily in all radio environments.

FUTURE TRENDS: A UNIFIED FRAMEWORK FOR FACILITATING MULTI-FUNCTIONED MOBILITY OVER HETEROGENEOUS NETWORKS

It should become clear from the solutions presented previously that the problem of integrating mobility, quality of service, and security into a single network access platform often has been assumed to comprise three separate problem spaces that require the lateral interactions between them to be figured out and implemented. However, as we scrutinize more deeply into the problem, we see that the mobility function alone may comprise functions such as Mobile IP, movement detection, candidate access point/router discovery, smart interface selection, fast handover, seamless handover, handover target selection, context transfer, and so on. Furthermore, for the QoS function, based on recent developments, it may be desirable to have some sort of service class negotiation function, access control function, and a function to exact queue and L2-specific configurations. Similarly, for the security portion, at least an access authentication function, an access control function, and some function to secure the signaling transport and, optionally, the data transport may be needed. Moreover, if we need our software to work across more than one access technology in a performance-optimized manner, we may probably need different sets of functions per access technology. Such a requirement inevitably raises difficulties among designers of multi-functioned network-access protocols and platforms for future mobile Internet access over heterogeneous networks, as can be discerned from current discussions in the IETF Seamoby (SEAMOBY, 2003)and IETF Mipshop (MIPSHOP, 2003) working groups. In the current state of the art, complexities up to the order of N^2 regarding interworkings need to be solved between N functions; up to N decisions and possible modifications with regard to interworking are further needed to integrate a new, (N+1)th function.

A unified framework, therefore, is required to overcome this complexity. In this framework, one can define a single, common interface that makes it possible for a multi-functioned network access platform supporting mobility to be decomposed into and treated as independent functions or protocols that can be separately designed, analysed, developed, integrated, tested, and deployed with the full system.

CONCLUSION

In recent years, there has been much interest in implementing network access protocols and platforms that integrate mobility, QoS, and security related functions over heterogeneous Internet access networks. In this article, we presented a comprehensive overview of the various IP mobility solutions using leading examples of state-of-the-art research and development solutions, as well as solutions available commercially. We concluded the article by highlighting the need to have a unified framework that will resolve the potential complexity in providing a multi-functioned network access platform supporting mobility.

REFERENCES

Calhoun, P., & Perkins, C. (2000a). Mobile IP network access identifier extension for IPv4. *RFC 2794*.

Calhoun, P., & Perkins, C. (2000b). Mobile IP foreign agent challenge/response extension. *RFC 3012*.

Deering, S., & Hinden, R. (1998). Internet protocol, version 6 (IPv6) specification. *RFC 2460*.

Droms, R. (Ed.) (2003, July). Dynamic host configuration protocol for IPv6 (DHCPv6). *RFC3315*.

Einsiedler, H., et al. (2001). The Moby Dick project: A mobile heterogeneous ALL-IP architecture. *Proceedings of Advanced Technologies, Applications and Market Strategies for 3G ATAMS 2001*, Kraków, Poland.

Forsberg, D., et al. (2003). Protocol for carrying authentication for network access (PANA). Draft-ietf-pana-pana-02, *IETF*.

Huitema, C. (2001). Shipworm: Tunneling IPv6 over UDP through NATs. Draft-ietf-ngtrans-shipworm-03.txt, *IETF*.

IBM's Everyplace Wireless Gateway. Retrieved from *http://www.ibm.com*

Jayabal, R.J., et al. (2004). AMASE: An architecture for seamless IPv6 roaming & handovers with authentication & QoS provisioning over heterogeneous wireless networks [Internal technical report available from the authors April 2004].

Johnson, D., Perkins, C., & Arkko, J. (2003). Mobility support in IPv6. Draft-ietf-mobileip-ipv6-24.txt, *IETF*.

Kenward, G. (Ed.) (2002). General requirements for a context transfer. Draft-ietf-seamoby-ct-reqs-05.txt, *IETF*.

Koodli, R. (Ed.) (2003). Fast handovers for mobile IPv6. Draft-ietf-mipshop-fast-mipv6-00.txt, *IETF*.

Levkowetz, H., et al. (2003). Mobile IP traversal of network address translation (NAT) devices. *Request for Comment Documents (RFC 3519), Internet Engineering Task Force (IETF)*. Retrieved from *http://www.ietf.org*

Loughney, J. (Ed.) (2003). Context transfer protocol. Draft-ietf-seamoby-ctp-05.txt, *IETF*. [Work in progress].

MIPSHOP. (2003). IETF MIPv6 signaling and handoff optimization (mipshop) working group. *IETF*. Retrieved from *http://www.ietf.org/html.charters/mipshop-charter.html*

Moby Dick. (2003). Moby Dick: Mobility and differentiated services in a future IP network. Retrieved December 2003 from *http://www.ist-mobydick.org/*

Montenegro, G. (2002). Reverse tunneling for mobile IP, revised. *Request for Comment Documents (RFC 3024)*. *Internet Engineering Task Force (IETF).* Retrieved from *http://www.ietf.org*

NetMotion Wireless. (n.d.). Retrieved from *http://www.netmo tionwireless.com/*

PacketAir Networks. (n.d.). Retrieved from *http://www.packetair.com/*

Perkins, C. (1996). IP encapsulation within IP. *Request for Comment Documents (RFC 2003), Internet Engineering Task Force (IETF)*. Retrieved from *http://www.ietf.org*

Perkins, C. (2002). IP mobility support for IPv4. *Request for Comment Documents (RFC 3344). Internet Engineering Task Force (IETF)*. Retrieved from *http://www.ietf.org*

Rigney, C. et al. (2000). Remote authentication dial in user service (RADIUS). *RFC2865, Internet RFC, IETF*.

SEAMOBY. (2003). IETF context transfer, handoff candidate discovery, and dormant mode host alerting (Seamoby) working group. Retrieved from *http://www.ietf.org/html.charters/seamoby-charter.html*

Thomson, S., & Narten, T. (1998). IPv6 stateless address autoconfiguration. *RFC 2462, Internet RFC, IETF*.

KEY TERMS

Application Layer: Layer 7 of the OSI model. This layer determines the interface of the system with the user.

Bandwidth: The difference in Hertz between the limiting (upper and lower) frequencies of a spectrum.

Broadband: In data communications, generally refers to systems that provides user data rates of greater than 2 Mbps and up to 100s of Mbps.

Cellular Network: A wireless communications network in which fixed antennas are arranged in a hexagonal pattern, and mobile stations communicate through nearby fixed antennas.

Encapsulation: The addition of control information by a protocol entity to data obtained from a protocol user.

Network Layer: Layer 3 of the OSI model. Responsible for routing data through a communication network.

Protocol: A set of rules governing the exchange of data between two entities.

Protocol Data Unit: A set of data specified in a protocol of a given layer, consisting of protocol control information and possibly user data of that layer.

Router: A device used to link two or more networks. The router makes use of an Internet protocol, which is a connectionless protocol operating at layer 3 of the OSI model.

Service Access Point: A means of identifying a user of the services of a protocol entity. A protocol entity provides one or more SAPs for use of higher-level entities.

Session Layer: Layer 5 of the OSI model. Manages a logical connection (session) between two communicating processes or applications.

Transport Layer: Layer 4 of the OSI model. Provides reliable, transparent transfer of data between end points.

Wireless: Refers to transmission through air, vacuum, or water by means of an antenna.

Modeling Interactive Distributed Multimedia Applications

Sheng-Uei Guan
National University of Singapore, Singapore

INTRODUCTION

In recent years, researchers have tried to extend Petri net to model multimedia. The focus of the research flows from the synchronization of multimedia without user interactions, to interactions in distributed environments (Bastian, 1994; Blakowski, 1996; Diaz, 1993; Guan, 1998; Huang, 1998; Huang, 1996; Little, 1990; Nicolaou, 1990; Prabhakaran, 1993; Prabhat, 1996; Qazi, 1993; Woo, 1994). The issues that concern us are the flexibility and compactness of the model. Petri net extensions have been developed to facilitate user interactions (UI) in distributed environments; however, they require sophisticated pre-planning to lay out detailed schedule changes. In this article, we introduce a Reconfigurable Petri Net (RPN).

An RPN is comprised of a novel mechanism called a modifier (f), which can modify an existing mechanism (e.g., arc, place, token, transition, etc.) of the net. A modifier embraces controllability and programmability into the Petri net and enhances the real-time adaptive modeling power. This development allows an RPN to have a greater modeling power over other extended Petri nets. The article introduces both the model and theory for RPN and a simulation to show that RPN is feasible.

BACKGROUND

Little (1990) has proposed the use of Object Composition Petri Net (OCPN) to model temporal relations between media data in multimedia presentation. The OCPN model has a good expressive power for temporal synchronization. However, it lacks power to deal with user interactions and distributed environments. Extended Object Composition Petri Net (XOCPN), proposed by Woo, Qazi, and Ghafoor (1993), is an improved version of OCPN with the power to model distributed applications, but it does not handle user interactions.

The lack of power in OCPN to deal with user interactions has led to the development of an enhanced OCPN model, Dynamic Timed Petri Net (DTPN) proposed by Prabhakaran and Raghavan (1993). DTPN provides the ability for users to activate operations like skip, reverse, freeze, restart, and scaling the speed of presentation.

Guan (1998) has proposed DOCPN to overcome the limitations of the original OCPN and XOCPN. DOCPN extends OCPN to a distributed environment using a new mechanism known as prioritized Petri nets (P-nets) together with global clock and user interaction control. Guan and Lim (2002) later proposed another extended Petri net: Enhanced Prioritized Petri Net (EP-net), an upgraded version of P-net. It has a Premature/Late Arriving Token Handler (PLATH) to handle late and/or premature tokens (locked tokens forced to unlock). Moreover, EP-net has another feature: a dynamic arc that simplifies and improves the flexibility of designing interactive systems.

None of the above-mentioned Petri nets have controllability and programmability built in as RPN has offered to Petri net, neither do they have the ability to model a presentation on the fly and simulate real-time adaptive application.

RECONFIGURABLE PETRI NET

Definitions

RPN consists of two entities: control and presentation layers. Each entity is represented as a rectangle. These two layers can be joined together by a link (denoted by a double line). A link represents necessary interactions between the control layer and the presentation layer. Note that multiple presentation

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. RPN: A simple example

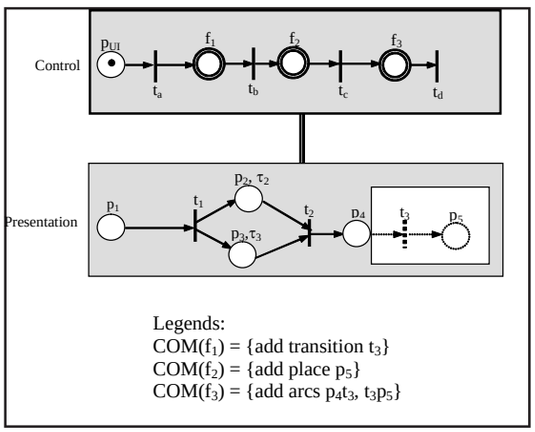
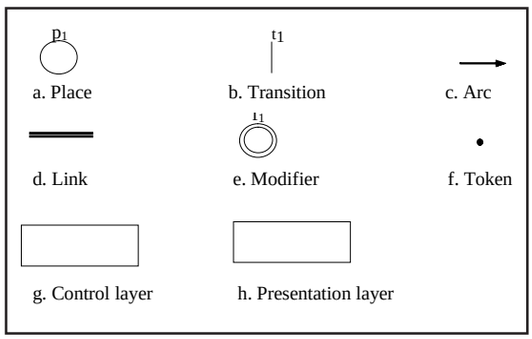


Figure 2. RPN graphic representations



and control layers could exist in a model. An example of RPN is shown in Figure 1. Initially, there are no mechanisms (e.g., t_3 and p_5) inside the white box of the presentation layer. After the activation of the modifiers (e.g., f_1, f_2, and f_3) in the control layer, the mechanisms inside the white box are created. First, the control layer as shown in Figure 1 starts with a token in p_{UI}, transition t_a is enabled and fires. The token is removed from p_{UI} and created at modifier f_1. Upon the token arriving at modifier f_1, transition t_3 is then created in the presentation layer. Transition t_b is enabled and fires only if a token is present at f_1 and the transition t_3 is created. After transition t_b is enabled and fires, the token in modifier f_1 is removed and created at modifier f_2. Upon the token arriving at modifier f_2, the place p_5 is created in the presentation layer. Next, transition t_c is enabled and fires. The token in modifier f_2 is removed and created at modifier f_3. Upon the token arriving at modifier f_3, the arcs p_4t_3 and t_3p_5 are created in the presentation layer. Finally, transition t_d is enabled and fires. The token is removed from modifier f_3. Therefore, we have shown how a RPN works. In the following, we explain the definitions related to RPN.

Definition: Control and Presentation Layers

The structure of an unmarked layer in RPN is a six-tuple, S = {T, P, A, D, L, *COM*}. For marked RPN's layer, the definition of the structure becomes a seven-tuple, S = {T, P, A, D, L, *COM*, M}. Refer to the structures S as mentioned above, where $P \cap T = \varnothing$. A complete RPN net may consist of zero or more control layers and one or more presentation layers.

$T = \{t_1, t_2, t_3, \ldots\ldots\ldots\ldots, t_m\}$ is a finite set of transitions where $m > 0$.

$P = \{p_1, p_2, p_3, \ldots\ldots\ldots, p_i, f_4, f_5, f_6, \ldots\ldots\ldots f_k\}$ is a finite set of places and/or modifiers where i and $k > 0$.

$COM: f_a \rightarrow \{com_1, com_2, com_3 \ldots\ldots com_z\}$ is a mapping from the set of modifiers to the commands (as defined in Table 1) where a and $z > 0$.

$A: \{P \times T\} \cup \{T \times P\}$ is a set of arcs representing the flow relation.

$M: P \rightarrow I^+$, $I^+ = \{0, 1, 2\ldots\}$ is a mapping from the set of places or modifiers to the integer numbers, representing a marking of a net.

$D: p_b \rightarrow R^+$ is a mapping from the set of places to the non-negative real numbers, representing the presentation intervals or the durations for the resources concerned where $b > 0$.

$L = \{c_x$ or $p_x\}$ indicates whether an entity is a control layer c_x or presentation layer p_x where $x > 0$.

The set of graphical symbols for RPN are demonstrated in Figures 2a to 2h. A classic place is shown in Figure 2a, which represents a resource (e.g., audio, video playback, etc.). If the place is associated with duration (D), it indicates the interval of the resource to be consumed. Figure 2b displays a transition, which represents a synchronization point in a presentation. In Figure 2c, an arc is demonstrated which represents a flowing relation in a presentation. Then, the links as shown in Figure 2d establish connections linking two

different layers (e.g., between control and presentation/control layers). Figure 2e introduces a modifier, which signifies a place having the ability to control, create, or delete a new or existing mechanism (e.g., arc, place, token, and transition) of a presentation/control layer in a net. A solid dot (token) as displayed in Figure 2f indicates the marking in a place. Finally, the two rectangles denote a control and a presentation layer as presented in Figures 2g and 2h respectively.

Definition 2: Firing Rules

A transition is enabled when all input places that are connected to it via an input arc have at least one token. If the condition mentioned is met, the transition fires and token(s) are removed from each of its input places and token(s) is created at each of its output places. The transition fires instantly if each of its input places contains an unlocked token. In case when a place is associated with duration, the place remains in the active state for an interval specified by the duration d_1 after receiving a token. During this period, the token is locked. At the end of duration d_1, the token becomes unlocked.

RPN extends the capabilities of OCPN by providing support for interactive distributed multimedia environment and enhance the modeling power over the latest extended Petri net (e.g., P-net, EP-net, etc.). This is achieved by using some novel mechanisms: modifiers. The entities of a RPN are grouped into two layers: control and presentation layers. With mechanisms grouped into layers, modifiers can modify them in terms of group instead of individual. In a way, this helps reducing the size of modeling task. However, the grouping approach is not the key factor to reduce the size of modeling task. The key factor is the power introduced by the programming modifiers. Once the mechanisms are grouped into layers, links indicate the communication between layers.

Synchronization

Extended Petri nets are so popular in modeling multimedia presentation because they exhibit the

Table 1. List of commands

No.	Mechanisms	Commands	Actions
1.	Arc	Disable arc	An arc is disabled (virtually deleted).
		Enable arc	An arc is enabled (recovered).
		Create arc	An arc is created.
		Delete arc	An arc is deleted.
		Reverse arc	The direction of an arc is reversed.
2.	Place or Modifier	Create place or modifier	A place or modifier is created.
		Delete place or modifier	A place or modifier is deleted.
		Replace place or modifier	A place or modifier is replaced.
3.	Transition	Disable transition	A transition is disabled (virtually deleted).
		Enable transition	A transition is enabled (recovered).
		Create transition	A transition is created.
		Delete transition	A transition is deleted.
4.	Token	Lock token	To lock a token. (The duration continues to count down, however when the count reaches zero, the token remains lock)
		Unlock token	To unlock a token. (The duration forces to zero and the token is unlocked)
		Pause token	To stop counting down if a place associated with a duration or stops a transition to be fired if the place is associated with no duration.
		Resume token	To resume a token and start counting from the time it has been paused.
		Create token	To create a token to the indicated place with no condition.
		Delete token	To remove a token at the indicated place with no condition.

synchronization properties among resources, for example lip-sync. In this section, the term synchronization refers to the synchronization between layers. In order to prevent any conflict between the layers, a control layer should pause the token(s) in the presentation before carrying out its necessary executions. In case of controlling the presentation on fly, the user needs to anticipate outcome of his design to avoid any adverse result.

SYNCHRONOUS CONTROL OF USER INTERACTIONS

The synchronization mechanism (in the control layer): a modifier has the authority to manipulate the existing mechanisms or generate new mechanisms (in presentation layers). This enhances the power to support user interaction. Whenever a user interacts, a token arrives at the initial place p_{UI}. As the token flows through the RPN structure in the control layer, the modifiers with each associated commands are executed respectively and the interactions are carried out properly.

RPN provides the ability for users to activate operations like reverse, skip, freeze, and restart, and speed scaling operations. The reverse operation is similar to the forward operation, only that the presentation flows are opposite. Sometime, the reverse operation can also be combined with the speed scaling operation to form a fast reverse operation. A user might feel that a certain section of a presentation boring and decides to skip. This operation is able to skip on the fly an ongoing stage. Among various user interactions, freeze and restart operations are the simplest ones to model. In speed scaling, a user can either increase or decrease the speed of a presentation by a factor of 2, 3, and so on in the forward or reverse manner.

RPN has the ability of reducing the size of modeling task as compared to other extended Petri nets. To model a lip-sync presentation of a series of video frames and audio samples, for example 1000 frames or samples based on existing extended Petri nets, the user needs to create about 2000 places (representing the video frames and audio samples). As a result, this has become an intricate and time-wasting task for the user.

IMPLEMENTATION: RPN SIMULATION

The modifiers and layers (i.e. control and presentation layers) of RPN are introduced to enable users to specify user interactions. The presentation-specification mechanisms like places, input/output arcs, and transitions are developed to enable users to specify temporal relationships among media in a presentation. Together, the distributed interactive multimedia applications can be simulated using this RPN simulator. These mechanisms might be grouped into many different presentation layers, whereby some of these layers might be monitored and manipulated by other control layers. This is an interesting issue because we have formed an object-oriented approach to the model. The control layer uses to control a presentation layer, can also use to control other layers in future.

RPN simulator is designed to be user-friendly. What a user needs is a mouse that does most of the job. To draw a place, modifier, or transition, the user just clicks on the place, modifier, or transition icon shown on top of the menu as displayed in Figure 3, and keys in an integer label from 1 to 50. In the current prototype, we have set its maximum label to 50. Then, by clicking onto any area outside the layers (see Figure 3), a place, modifier, or transition will be drawn. With that, the places or modifiers and transitions can be linked together with those event mechanisms by clicking the icons such as input event or output event show on top of the menu. Then the mechanisms are grouped together according to the control and presentation layers as shown in Figure 3. After the marking is initialized, the simulator is ready to run. This simulator has two running modes. The first mode runs step by step, which means it fires all enabled transitions once and waits for the next execution. The second mode runs and fires till no transition is enabled. The simulator simulates an example: mute operation during an MTV playback as demonstrated in Figure 3.

Each place has a local timer. The timer is initialized to a duration value when the presentation starts. The runtime executive in the simulator periodically updates the timer value associated with each active place. For example, places, p_2, p_3, p_4, and p_5 are associated with duration five seconds, place p_7 is associated with duration two seconds and places p_1

Figure 3. RPN Simulator: Mute operation during a mini MTV playback

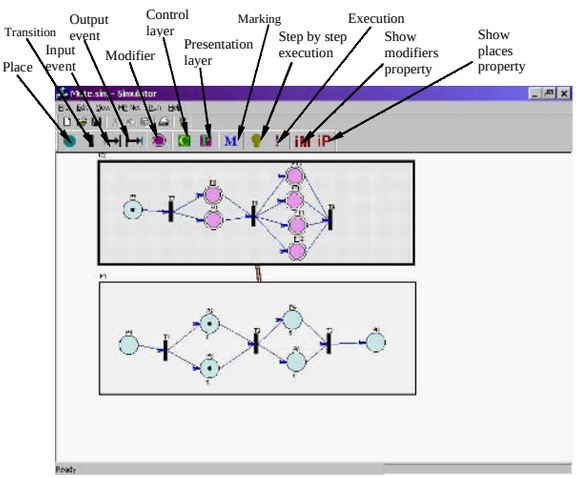


Figure 4. Dialog box of modifiers (see Figure 3) property

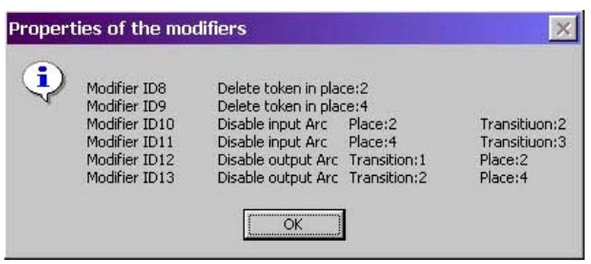

Properties of the modifiers
Modifier ID8 Delete token in place:2
Modifier ID9 Delete token in place:4
Modifier ID10 Disable input Arc Place:2 Transitiuon:2
Modifier ID11 Disable input Arc Place:4 Transitiuon:3
Modifier ID12 Disable output Arc Transition:1 Place:2
Modifier ID13 Disable output Arc Transition:2 Place:4
OK

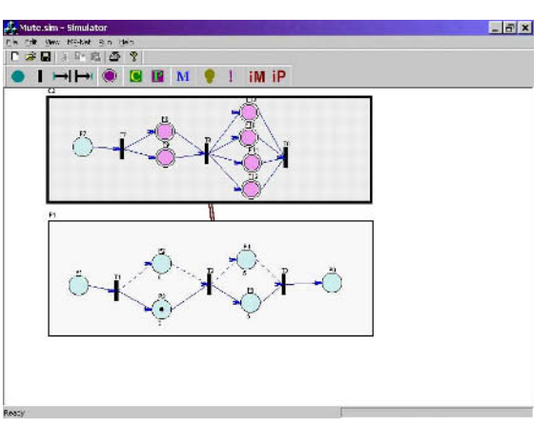
Figure 5. Dialog box of places (see Figure 3) property

and p_6 have not. If any of these places contain a token (locked), and the user runs the simulator, the duration will count down until zero. Upon the duration reaching zero, the token in the place is unlocked and ready to be removed if its transition fires. On the other hand, the token in the modifier is ready to be removed only after its command is executed and then if its transition fires.

Figures 4 and 5 show what happens when the icons "iM" and "iP" respectively are clicked. A dialog box pops up on the screen and this box indicates the legend of the modifiers or places as illustrated in Figure 3. In Figure 4, it shows that modifiers f_8 and f_9 are commanded to delete tokens contained in places p_2 and p_4, modifiers f_{10} and f_{11} are commanded to disable the input arcs p_2t_2 and p_4t_3 and modifiers f_{12} and f_{13} are commanded to disable the output arcs t_1p_2 and t_2p_4. The legend of the places in Figure 3 is self-explained by Figure 5.

Once the user clicks on the icon "!" (execution) on top of the menu, the simulation starts to run. Figure 6 shows a snapshot of the simulation three seconds later after the user initialized the simulator to run. At this instant, the modifiers have executed their commands. Therefore, the programmed tokens, input arcs, and output arcs are deleted and disabled respectively. In other words, the model has simulated a mute mode of a mini MTV playback.

Figure 6. RPN simulation: After the modifiers executed their commands



CONCLUSION AND FUTURE WORK

We have proposed a powerful synchronization mechanism RPN for multimedia synchronization control where schedule changes can be made into the presentation layer at run-time. With the comprehensive commands that can be associated with a modifier, the modeling power of RPN is much greater than the conventional Petri net and its extensions in terms of modeling user interactions. Some of the basic user interactions such as reverse, skip, freeze and restart, and speed scaling are modeled.

RPN facilitates the compact and flexible specification of run-time, large-scale specifications while preserving fine granularity as well as supporting real-time user interactions in distributed environment.

REFERENCES

Andleigh, P.K. & Thakrar, K. (1996). Multimedia systems design. *Prentice Hall PTR*, 421-444.

Bastian, F. & Lenders, P. (1994). Media Synchronization on Distributed Multimedia Systems. *International Conference on Multimedia Computing and System,* 526-531.

Blakowski, G. & Steinmetz, R. (1996). A media synchronization survey: Reference model, specification, and case studies. *IEEE Journal on Selected Areas in Communication,* (1), 5-35.

Diaz, M. & Senac, P. (1993). Time Stream Petri Nets A Model for Multimedia Streams Synchronization. *Proceedings of the First International Conference on Multimedia Modeling,* 257-273.

Guan, S., Hsiao-Yeh, Y. & Jen-Shun, Y. (1998). A prioritized Petri net model and its application in distributed multimedia systems. *IEEE Transactions on Computers,* (4), 477-481.

Guan, S. & Lim, S. (2002). Modeling multimedia with enhanced prioritized Petri nets. *Computer Communications,* (8), 812-824.

Huang, C. & Lo, C. (1996). An EFSM-based multimedia synchronization model and the authoring system. *IEEE Journal on Selected Areas in Communication,* (1), 138-152.

Huang, C. & Lo, C. (1998). Synchronization for Interactive Multimedia Presentations. *IEEE Multimedia,*(4), 44-62.

Little, T. (1990). Synchronization and storage models for multimedia objects. *IEEE Journal on Selected Areas in Communication,* (3), 413-427.

Nicolaou, C. (1990). An architecture for real-time multimedia communication systems. *IEEE Journal on Selected Areas in Communications,*(3), 391-400.

Peterson, J. (1981). *Petri net theory and the modeling of systems.* New Jersey: Prentice-Hall, Inc.

Prabhakaran, B. & Raghavan, S.V. (1993). Synchronization Models for Multimedia Presentation with User Participation. *ACM Multimedia Proceedings*, 157-166.

Qazi, N., Woo, M. & Ghafoor, A. (1993). A Synchronization and Communication Model for Distributed Multimedia Objects. *Proceedings First ACM International Conference on Multimedia*, 147-155.

Woo, M., Qazi, N.U. & Ghafoor, A. (1994). A Synchronization Framework for Communication of Pre-orchestrated Multimedia Information. *IEEE Network*, (8), 52-61.

KEY TERMS

Distributed Environment: An environment in which different components and objects comprising an application can be located on different computers connected to a network.

Modeling: The act of representing something (usually on a smaller scale).

Multimedia: The use of computers to present text, graphics, video, animation, and sound in an integrated way.

Petri Nets: A directed, bipartite graph in which nodes are either "places" (represented by circles) or "transitions" (represented by rectangles), invented by Carl Adam Petri. A Petri net is marked by placing "tokens" on places. When all the places with arcs to a transition (its input places) have a token, the

transition “fires”, removing a token from each input place and adding a token to each place pointed to by the transition (its output places). Petri nets are used to model concurrent systems, particularly network protocols.

Synchronization: In multimedia, synchronization is the act of coordinating different media to occur or recur at the same time.

Tokens: An abstract concept passed between places to ensure synchronized access to a shared resource in a distributed environment.

User Interaction: In multimedia, the act of users intervening or influencing in designing multimedia presentation.

Modelling eCRM Systems with the Unified Modelling Language

Călin Gurău
Centre d'Etudes et de Recherche sur les Organisations et la Management (CEROM), France

INTRODUCTION

Electronic commerce requires the redefinition of the firm's relationships with partners, suppliers, and customers. The goal of effective Customer Relationship Management (CRM) practice is to increase the firm's customer equity, which is defined by the quality, quantity, and duration of customer relationships (Fjermestad & Romano, 2003). The proliferation of electronic devices in the business environment has determined the companies to implement electronic customer relationship management (eCRM) systems, which are using advanced technology to enhance customer relationship management practices.

The successful implementation of an eCRM system requires a specific bundle of IT applications that support the following classic domains of the CRM concept: marketing, sales, and service (Muther, 2001). Electronic marketing aims at acquiring new customers and moving existing customers to further purchases. Electronic sales try to simplify the buying process and to provide superior customer support. Electronic service has the task of providing electronic information and services for arising questions and problems or directing customers to the right contact person in the organization.

The eCRM system comprises a number of business processes, which are interlinked in the following logical succession:

- **Market Segmentation:** The collection of historical data, complemented with information provided by third parties (i.e., marketing research agencies), is segmented on the basis of customer life-time value (CLV) criteria, using data mining applications.
- **Capturing the Customer:** The potential customer is attracted to the Web site of the firm through targeted promotional messages diffused through various communication channels.
- **Customer Information Retrieval:** The information retrieval process either can be implicit or explicit. When implicit, the information retrieval process registers the Web behavior of customers using specialized software applications such as cookies. On the other hand, explicit information can be gathered through direct input of demographic data by the customer (using online registration forms or questionnaires). Often, these two categories of information are connected at database level.
- **Customer Profile Definition:** The customer information collected is analyzed in relation to the target market segments identified through data mining, and a particular customer profile is defined. The profile can be enriched with additional data (e.g., external information from marketing information providers). This combination creates a holistic view of the customer, its needs, wants, interests, and behavior (Pan & Lee, 2003).
- **Personalization of Firm-Customer Interaction:** The customer profile is used to identify the best customer management campaign (CMC), which is applied to personalize the company-customer online interaction.
- **Resource Management:** The company-customer transaction requires complex resource management operations, which are partially managed automatically through specialized IT applications such as Enterprise Resource Planning (ERP) or Supply Chain Management (SCM) and partly through the direct involvement and coordination of operational managers.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

BACKGROUND

The effective functioning of the eCRM system requires a gradual process of planning, design, and implementation, which can be greatly enhanced through business modeling. The selection of an appropriate business modelling language is essential for the successful implementation of the eCRM system and, consequently, for evaluating and improving its performance (Kotorov, 2002). The starting point for this selection is the following analysis of the specific characteristics and requirements of the eCRM system (Opdahl & Henderson-Sellers, 2004; Muther, 2001):

- eCRM is an Internet-based system; therefore, the modelling language should be able to represent Web processes and applications;
- The interactive nature of eCRM systems requires a clear representation of the interaction between customers and Web applications as well as between various business processes within the organization;
- eCRM systems are using multiple databases that interact with various software applications; the modelling language should support data modeling profiles and database representation;
- The necessity for resource planning and control requires a clear representation of each business process with its inputs, outputs, resources, and control mechanisms;
- The implementation and management of an eCRM system requires the long-term collaboration of various specialists such as business and operational managers, programmers, and Web designers, which are sometimes working from distant locations; the modeling language should provide a standard, intuitive representation of the eCRM system and business processes in order to facilitate cross-discipline interaction and collaboration;
- The complexity of the eCRM system requires a modelling language capable of presenting both the organizational and functional architecture at the level of system, process, software applications, and resources; this will facilitate a multi-user, multi-purpose use of the same business model, although the detail of representation might differ, depending on the required perspective.

The Unified Modeling Language (UML) is the notation presented in this article to support the business process modeling activity. The UML is well suited to the demands of the online environment. It has an object-oriented approach and was designed to support distributed, concurrent, and connected models (Gomaa, 2000; Rumbaugh, Jacobson, & Booch, 2004).

THE UNIFIED MODELLING LANGUAGE (UML)

UML was developed in 1995 by Grady Booch, Ivar Jacobson, and Jim Rumbaugh at Rational Corporation (Maciaszek, 2001; Rumbaugh et al., 2004), with contributions from other leading methodologists, software vendors, and users. Rational Corporation chose to develop UML as a standard through the Object Management Group (OMG). The resulting cooperative effort with numerous companies led to a specification adopted by OMG in 1997.

UML has a number of specific advantages:

1. **Simplicity of Notation:** The notation set is simple and intuitive.
2. **Standard:** The UML standard achieved through the OMG gives confidence to modellers that there is some control and consideration given to its development.
3. **Support:** A significant level of support is available to modellers using the UML:

    - Textbooks that describe the UML notation and consider specific application areas (Stevens & Pooley, 2000).
    - Papers in journals and publications/resources on the Internet spread knowledge of the UML (e.g., Rational Resource Center and UML Zone).
    - Software tools, often referred to as Computer Aided Software Engineering (CASE) tools, are available. These provide support for documentation of UML

diagrams such as Rational Rose, argoUML, Objects By Design, and Enterprise Modeller. Training courses are available that instruct in the use of the core notation as well as general modeling concepts and use of associated CASE tools.

4. **Uptake:** The UML notation has quickly gathered momentum. This is driven by the need for such notation, assisted by the support mechanisms identified previously. The more the UML is used, the wider the knowledge pool becomes, which leads to a wider dissemination of information concerning the benefits and pitfalls of its use.
5. **Methodologies:** The development of methods or methodologies that provide support and guidelines on how to use the UML in a particular situation is widespread. A prime example is the Rational Unified Process (Siau & Halpin, 2001).
6. **Extensible:** The UML has a number of standard extension mechanisms to make the notation flexible: stereotypes, tagged values, and constraints (Eriksson & Penker, 2000; Kulak & Guiney, 2003).
7. **Living Language:** It is important to recognize UML as a living language; the standard is constantly developing, although in a controlled manner. The OMG works with representatives from various companies to clarify and address problems in the UML specification as well as considering recommendations for extensions to the language.

The UML is used to model a broad range of systems (e.g., software systems, hardware systems, databases, real-time systems, and real-world organizations). By sharing a common notation across system and business boundaries, the business and system analysts can better communicate their needs, being able to build a system that effectively solves customers' problems.

In addition, UML is developing in three main directions that are of interest for this article:

- **Data Modelling:** One or more databases are a component of almost all e-business applications, including CRM. Coordinating programming languages and databases has long been a difficult problem in system development, because each used a different method to declare data structure, leading to subtle inconsistencies and difficulties in exchanging information among programs and databases. UML has begun to address this problem by introducing a data modelling profile, which includes an additional set of notations to capture the data modelling and database connectivity aspects of modeling (Naiburg & Maksimchuk, 2003; Siau & Halpin, 2001).
- **WWW System Modeling:** The development of businesses and systems for the WWW has led to an extension of UML for modelling Web-based systems. This capability is provided as a UML profile that enables modellers to represent various elements that compose a Web application (e.g., client pages, server pages, forms, frames, etc.). The profile contains a set of stereotypes for different elements and their relationships (Conallen, 2000).
- **Business Process Modelling:** Important extensions to UML concern notations suggested to fully describe the processes, goals, and rules of business (Eriksson & Penker, 2000).

USING UML TO REPRESENT ECRM SYSTEMS

Additional extensions to UML have been proposed to support business modelling (Kim, 2004). The Eriksson-Penker Business Extensions (Eriksson & Penker, 2000) adapt the basic UML activity diagram and introduce a so-called process diagram. Table 1 describes the notation used in this section.

An example of an Eriksson-Penker process diagram is shown in Figure 1. The diagram represents a process and the objects involved in that process. The process is triggered by an event and outputs a further resource. The use of the UML stereotype notation clarifies the role of each object (e.g., <<goal>>, <<information>>) and association (e.g., <<achieve>>, <<supply>>), as necessary. The direction of associations clearly shows the input and output relationship that objects have with the given process symbol.

Table 1. Summary of UML notation used in this article

Modelling icon	Name	UML Definition
<<text here>>	Stereotype	The text shown in chevron brackets is used for extra clarification.
<<process>> Name	Business process	A process, takes input resources from its left-hand side and indicates its output resources on its right-hand side (shown as dependencies to and from the process, according to standard UML syntax). The process symbol may also include the stereotype `<<process>>`, which is a textual description of the process.
Name	Business object	An object which is input to or output from an object. A stereotype may be added to clarify process goals (`<<goal>>`), physical resource (`<<resource>>`), or people (`<<people>>`).
<<information>> Name	Information object	An object, which is specifically identified as information. The alternative icon is used for clarity.
<<event>> Name	Event	An event is the receipt of some object, a time or date reached, a notification or some other trigger that initiates the business process. The event may be consumed and transformed (for example a customer order) or simply act as a catalyst.
---------->	Dependency	Connecting line with arrow shows dependencies between model components. Direction of arrow

Figure 1. Eriksson-Penker process diagram

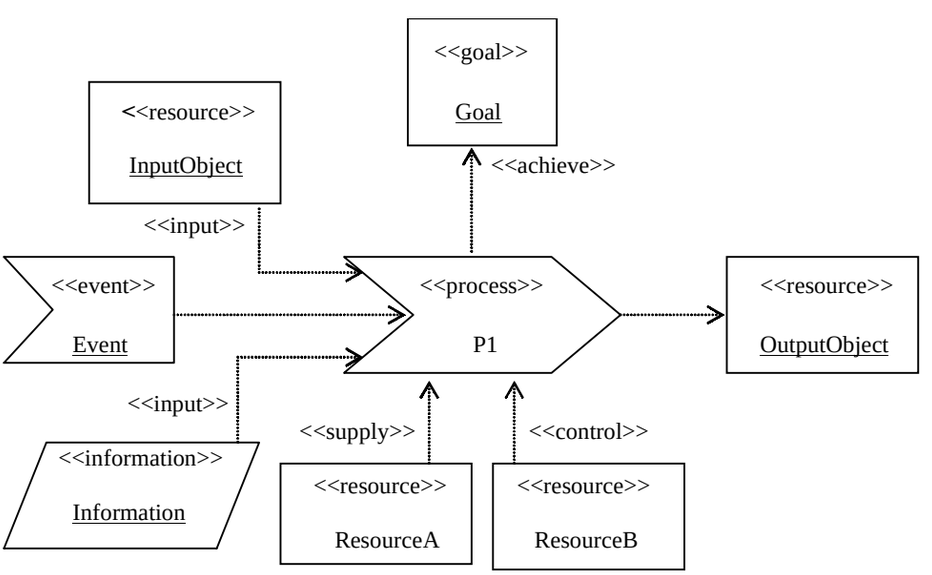
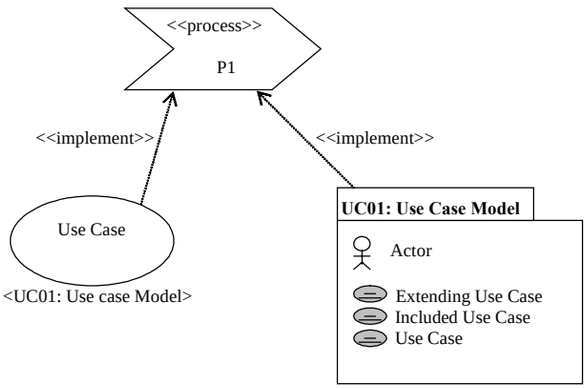


Figure 2. Example of implementation diagram



Using the Eriksson-Penker process diagram, the implementation process of an eCRM system will be further presented and analyzed. The process is common for every type of e-business, and the diagrams presented can be used as business modelling frameworks by any Internet-based organization. On the other hand, in order to keep the model simple and easy to understand, the diagrams only show the major business processes involved in the system. The development of these diagrams to include more specific and detailed processes can and must be done by every business organization, depending on its goals, structure and strategy.

The business process diagram also allows a detailed representation of the way in which a given business process is implemented in a system. Using an implementation diagram, use cases, packages, and other model artefacts may be linked back to the business process with <<implementation>> links to signify a dependent relationship (Kulak & Guiney, 2003). The example provided in Figure 2 illustrates how a business process is implemented by a use case and a package. As the model develops and the functional software components are built and linked to use cases, the business justification for each element can be derived from this representation.

To increase the accuracy of the representation, the model presented in Figure 2 also implies what is not being delivered. Since the business process typically will include a wide range of manual and automated procedures, this model illustrates exactly what functionality (use cases) needs to be provided to service a particular business process; on the other hand, any missing functionality must be outsourced from other systems and/or procedures.

Using UML notations, the main business processes involved in eCRM systems can be represented as follows.

eCRM Process 1: Segmenting the Market (Figure 3)

In order to segment the market, the firm needs to collect data about its customers. This can be done

Figure 3. Market segmentation

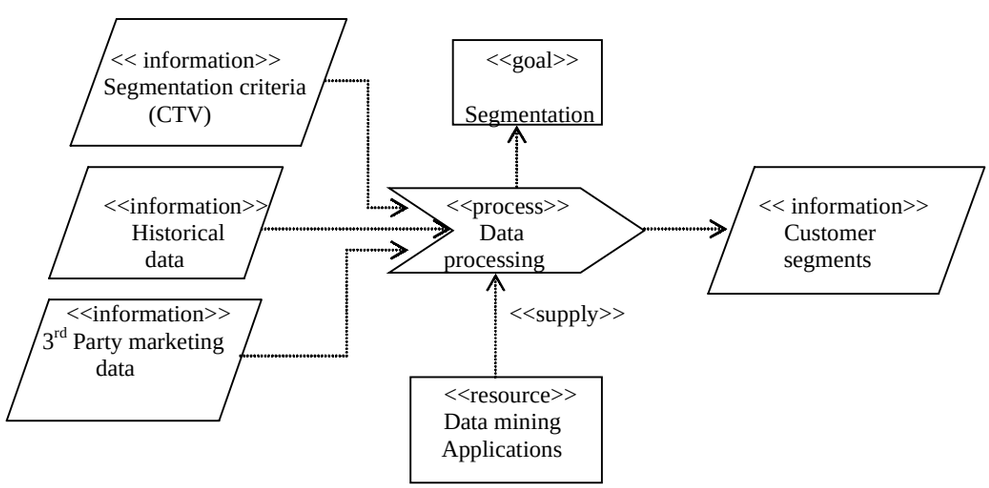
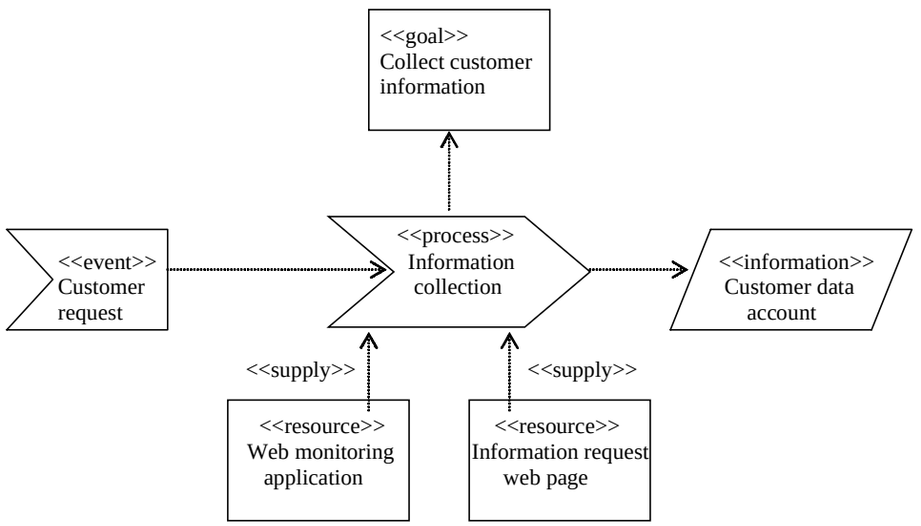


Figure 4. Customer information retrieval



either through online automated systems that register the history of customer-firm interaction (historical data) or by buying the necessary data from a third party (usually a specialized market research agency). These data will be usually located in databases. Applying the CLV method and using the segmentation criteria established by marketing managers, the collected data can be automatically processed using data mining applications such as pattern recognition and clustering. The output will represent a database of various customer segments that have different lifetime values (value segmentation) and, therefore, present different levels of priority for the firm (Rosset, Neumann, Eick & Vatnik, 2003; Wilson, Daniel & McDonald, 2002).

eCRM Process 2: Capturing the Customer

This process is not represented in this article, since it implies a multiple channel strategy and interaction. The customers can be attracted to the company's Web site either through promotional messages or through word-of-mouth referrals. The access to the

Figure 5. Customer profile definition

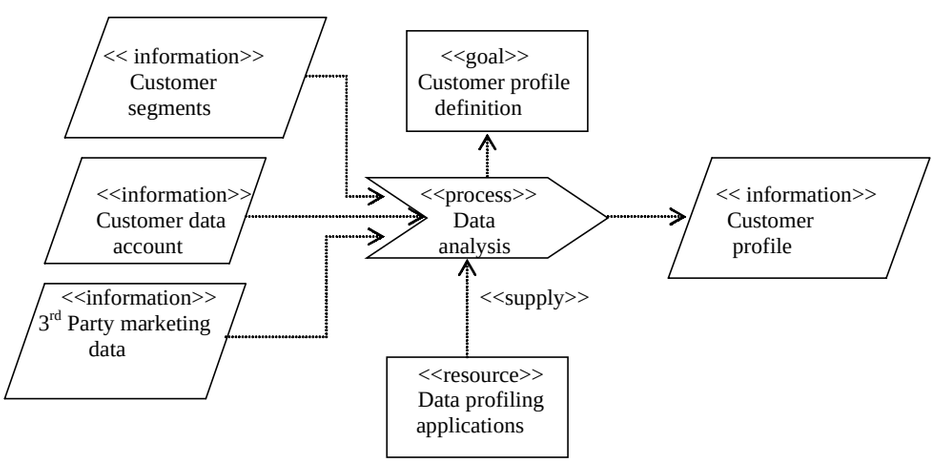
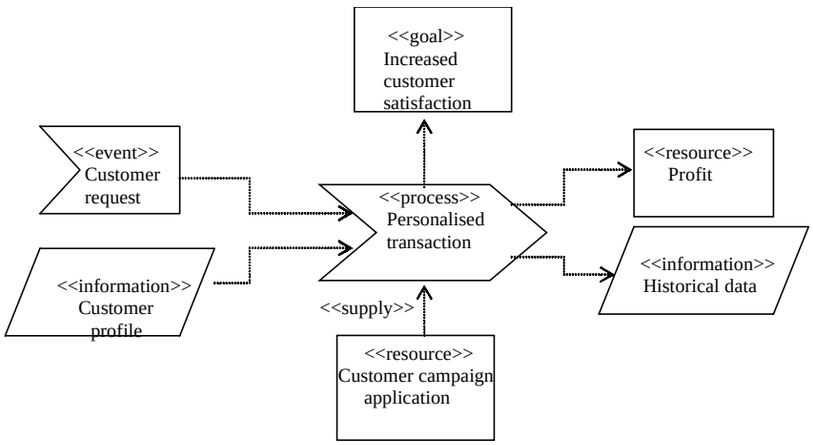


Figure 6. The personalization of customer-firm transaction



company Web site will be made using various intermediaries (i.e., search engines or company directories) and Web applications (i.e., hyperlinks).

eCRM Process 3: Customer Information Retrieval (Figure 4)

The customer information retrieval process usually will be initiated by the customer's request for a product or service (<<event>>). The information retrieval can be implicit (using Web-tracking applications) or explicit (using information request Web pages). The retrieved information is collected in a specific customer database account.

eCRM Process 4: Customer Profile Definition (Figure 5)

The information contained in the customer data account is analyzed and compared with the customer segments identified in the market segmentation stage, and a specific customer profile is defined. In order to refine this profile, additional information can be outsourced from specialized marketing agencies.

eCRM Process: Personalized Customer-Firm Transaction (Figure 6)

To increase the loyalty of the most profitable customers, the company needs to design and implement customized e-marketing strategies (Tan, Yen & Fang, 2002; Wilson et al., 2002).

The customer profile defined in the previous stage will be matched with the most effective customer campaign applications, determining the personalization of company-customer interactions. The completed transaction results in profits for the firm, increased satisfaction for customers, as well as information, which is integrated in the transaction history of that particular customer.

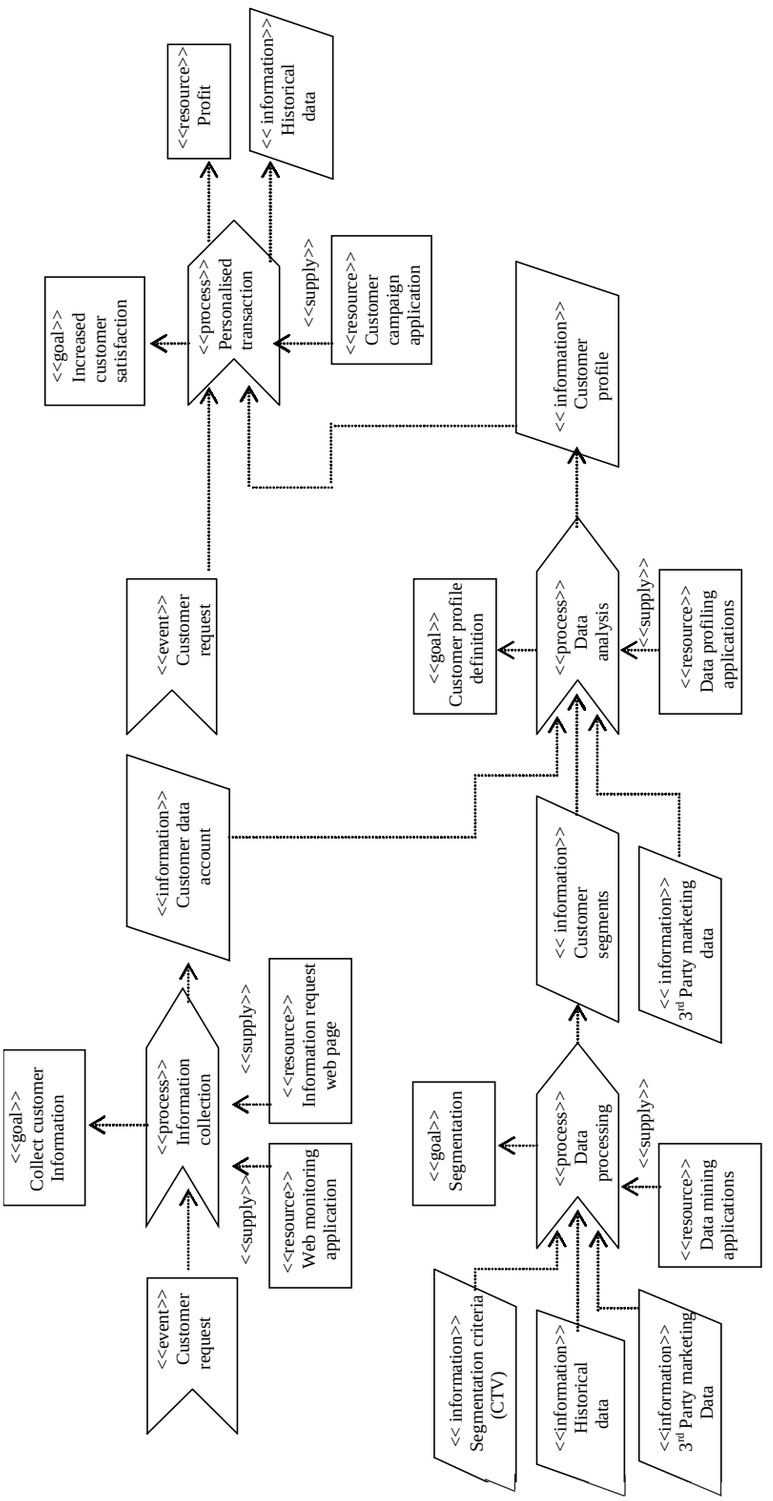
eCRM Process 6: Resource Management

This particular process involves complex interactions among operational managers, the company, and the firm's network of suppliers. The modelling of this business process requires advanced network modelling procedures. UML can be used efficiently to represent the networked interactions between the firm and external suppliers, being a distributed and highly standardized modeling language.

The Integration of Business Processes in the eCRM System (Figure 7)

Figure 7 presents four main business processes integrated into the eCRM system. The model shows how the outputs of one stage represent the inputs for the next stage. The resulting historical data at the

Figure 7. The integration of business processes in the eCRM system



end of the process closes the loop and restarts the process for a better tuning of the company's activities to the customers' needs.

Although only two of the represented business processes are visible to the online customer, the whole eCRM system uses software programs and applications, which either are Internet-based or are interacting closely with Web processes. Additional representation details can be included in the model, depending on the end-user orientation.

CONCLUSION

Because of its complexity, the successful implementation of an eCRM system requires a preliminary effort of business analysis, planning, and modelling. The choice of an appropriate modeling language is a necessary and essential step within this process.

This article attempted to present the manifold utility of the UML for business modeling, which is advocated by many authors:

1. UML can be used to represent the workflow processes within the organization and especially the flow of information, which is essential for online businesses (Lin, Yang & Pai, 2002).
2. UML offers a complete semantics for database design and can provide a powerful neutral platform for designing database architecture and data profiling, especially in the case of multi-user databases (Naiburg & Maksimchuk, 2003; Siau & Halpin, 2001).
3. UML can be used to represent the interaction between the digital company and different types of customers, helping the operational managers to identify the areas and activities of value creation and those of value destruction (Kim, 2004).
4. UML provides the basis for designing and implementing suitable information systems that support the business operations. The use of UML both for software description and for business modeling offers the possibility of mapping large sections of the business model directly into software objects (Booch, 2000; Maciaszek, 2001).
5. UML can provide a protocol neutral modeling language to design the interface between cooperating virtual organizations (Kotorov, 2002; Tan et al., 2002).
6. The capacity of the UML to provide a common platform for representing both Web processes and organizational architecture offers a unifying tool for the multi-disciplinary team that designs, implements, and controls the eCRM system (Siau & Halpin, 2001).

The business modeling exercise should be based on an analytical and modular approach. The implementation and functioning of the eCRM system must be represented stage-by-stage, taking into account, however, the final integration into a complete, functional system, as it was presented in this article.

Finally, it is important to understand the precise functions and limitations of modeling languages. The UML cannot guarantee the success of eCRM systems but establishes a consistent, standardized, tool-supported modeling language that provides a framework in which practitioners may focus on delivering value to customers.

REFERENCES

Booch, G. (2000). Unifying enterprise development teams with the UML. *Journal of Database Management, 11*(4), 37-40.

Conallen, J. (2000). *Building Web applications with UML*. London: Addison Wesley Longman.

Eriksson, H.-E., & Penker, M. (2000). *Business modelling with UML: Business patterns at work*. New York: John Wiley & Sons.

Fjermestad, J., & Romano Jr., N.C. (2003). Electronic customer relationship management. Revisiting the general principles of usability and resistance—An integrative implementation framework. *Business Process Management Journal, 9*(5), 572-591.

Gomaa, H. (2000). *Designing concurrent, distributed, and real-time applications with UML*. Reading, MA: Addison Wesley Object Technology Series.

Kim, H.-W. (2004). A process model for successful CRM system development. *Software IEEE, 21*(4), 22-28.

Kotorov, R.P. (2002). Ubiquitous organization: Organizational design for e-CRM business. *Process Management Journal, 8*(3), 218-232.

Kulak, D., & Guiney, E. (2003). *Use cases: Requirements in context*. Harlow, UK: Addison Wesley.

Lin, F.-R., Yang, M.-C., & Pai, Y.-H. (2002). A generic structure for business process modelling. *Business Process Management Journal, 8*(1), 19-41.

Maciaszek, L.A. (2001). *Requirements analysis and system design: Developing information systems with UML*. Harlow, UK: Addison-Wesley.

Muther, A. (2001). *Customer relationship management: Electronic customer care in the new economy*. Berlin: Springer-Verlag.

Naiburg, E.J., & Maksimchuk, R.A. (2003). UML for database design. *Online Information Review, 27*(1), 66-67.

Opdahl, A.L., & Henderson-Sellers, B. (2004). A template for defining enterprise modelling constructs. *Journal of Database Management, 15*(2), 39-74.

Pan, S.L., & Lee, J.-N. (2003). Using e-CRM for a unified view of the customer. *Communications of the ACM, 46*(4), 95-99.

Rosset, S., Neumann, E., Eick, U., & Vatnik, N. (2003). Customer lifetime value models for decision support. *Data Mining and Knowledge Discovery, 7*(3), 321-339.

Rumbaugh, J., Jacobson, I., & Booch, G. (2004). *Unified modelling language reference manual*. Harlow, UK: Addison-Wesley.

Siau, K., & Halpin, T. (2001). *Unified modelling language: Systems analysis, design and development issues*. Hershey, PA: Idea Group Publishing.

Stevens, P., & Pooley, R. (2000). *Using UML software engineering with object and components*. Harlow, UK: Pearson Education Limited.

Tan, X., Yen, D.C., & Fang, X. (2002). Internet integrated customer relationship management—A key success factor for companies in the e-commerce arena. *Journal of Computer Information Systems, 42*(3), 77-86.

Wilson, H., Daniel, E., & McDonald, M. (2002). Factors for success in customer relationship management (CRM) systems. *Journal of Marketing Management, 18*(1/2), 193-219.

KEY TERMS

Concurrent Models With an Object-Oriented Approach: Each object can potentially execute activities or procedures in parallel with all others.

Connected Modles With an Object Oriented Approach: Each object can send messages to others through links.

Constraints: Extensions to the semantics of a UML element. These allow the inclusion of rules that indicate permitted ranges or conditions on an element.

Customer Lifetime Value (CLV): Consists of taking into account the total financial contribution (i.e., revenues minus costs) of a customer over his or her entire life of a business relationship with the company.

Distributed Models With an Object-Oriented Approach: Each object maintains its own state and characteristics, distinct from all others.

Electronic Customer Relationship Management (eCRM): CRM comprises the methods, systems, and procedures that facilitate the interaction between the firm and its customers. The development of new technologies, especially the proliferation of self-service channels like the Web and WAP phones, has changed consumer buying behavior and forced companies to manage electronically the relationships with customers. The new CRM systems are using electronic devices and software applications that attempt to personalize and add value to customer-company interactions.

Eriksson-Penker Process Diagram: UML extension created to support business modelling, which adapts the basic UML activity diagram to represent business processes.

Stereotypes: Extensions to the UML vocabulary, allowing additional text descriptions to be applied to the notation. The stereotype is shown between chevron brackets <<>>.

Tagged Value: Extensions to the properties of a UML element.

Multimedia Communication Services on Digital TV Platforms

Zbigniew Hulicki
AGH University of Science and Technology, Poland

INTRODUCTION

Digital television (TV)-based communication systems provide cost-effective solutions and, in many cases, offer capabilities difficult to obtain by other technologies (Elbert, 1997). Hence, many books and papers on digital TV have been published in recent years (Burnett, 2004; Collins, 2001; Dreazen, 2002; ETR, 1996; Mauthe, 2004; Scalise, 1999; Seffah, 2004; Whitaker, 2003). None of them, however, provide an exhaustive analysis of the service provision aspects at the application layer. Therefore, this contribution aims to fill that gap, with a comprehensive view on the provision of services on DTV platform.

MULTIMEDIA SERVICES ON TV PLATFORM

Digital video broadcasting (DVB) is a technology readily adaptable to meet both expected and unexpected user demands (DVB, 1996; Raghavan, 1998), and one can use it for providing bouquets of various services (Fontaine, 1997; Hulicki, 2001). Because it is still unclear exactly which multimedia services will be introduced and how the advent of digital technology alters the definition of the audio-visual media and telecoms markets and affects the introduction of new services, one has to consider a number of various aspects and issues dealing with the definition, creation and delivery of DTV services. Under consideration will be also a question of the possible substitutions of products and services which, previously, were not substitutable, and now result in new forms of competition.

Digital Multimedia TV Services

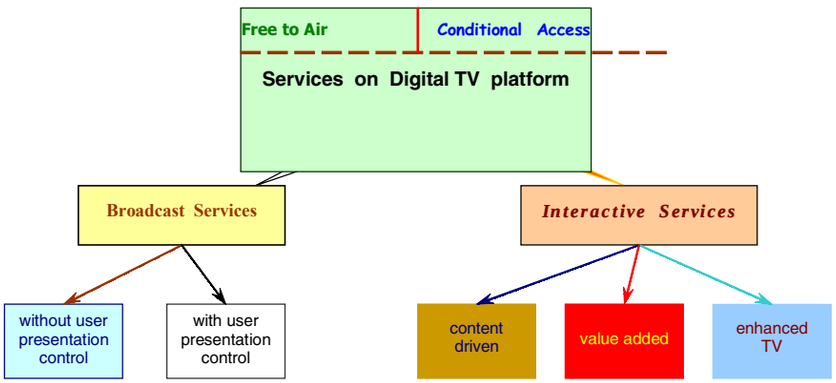
The advantage of the DTV (DTV) platform is the ability to provide a rich palette of various services, including multimedia and interactive applications, instead of providing only traditional broadcast TV services (Hulicki, 2000). To explore different services that can be provided via DTV systems, a generic services model is to be defined. This model will combine types of information flows in the communication process with categorization of services.

Depending on different communication forms and their application, two categories of telecommunications services can be distinguished on DTV platform, broadcast (or distribution) and interactive services (see Figure 1). These categories can be further divided into several subcategories (de Bruin, 1999); that is, the distribution subcategory will include services with and without individual user presentation control, while the registration, conversational, messaging and retrieval services will constitute a subcategory of the interactive services.

The interactive services will be the most complex, because of numerous offerings and a widely differing range of services with flexibility in billing and payment (Fontaine, 1997°). However, based on the object and content of services, some of them will refer to multimedia services, whereas the others will continue to be plain telecommunication services (see Figure 2). On the other hand, depending on the content's economic value, some of these services may be provided via conditional access (CA) system and will constitute the category of conditional access services. A CA system ensures that only users with an authorized contract can select, receive, decrypt and watch a particular TV programming package (EBU, 1995;

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. TV services categorized according to the form of communication



Lotspiech, 2002; Rodriguez, 2001). None of the networks currently in operation gives the possibility of providing all these services, but DTV seems to have a big potential for this (Hulicki, 2002).

The traditional principle of analog television is that the broadcaster's content is distributed via broadcast network to the end user and, with respect to these kinds of services, television can be considered a passive medium. Unlike analog, DTV enables more than the distribution of content only; that is, it allows a provision of interactive multimedia services. This implies that a user is able to control and influence the subjects of communication via interactive network (ETSI, 2000) (see Figure 3). Even though the user is able to play a more active role than before, the demand for interactive multimedia services continues to be unpredictable. Nevertheless, as the transport infrastructure is no longer service dependent, it becomes possible to integrate all services and evolve gradually towards interactive multimedia (Tatipamula, 1998; Raghavan, 1998).

The functions required for service distribution are variable and can be addressed in accordance with three main parameters: bandwidth, interactivity and subscriber management. The services to be developed will have variable transmission band requirements according to the nature of information transmitted (voice, data, video), the quality of the transmitted image and the compression techniques employed (Furht, 1999). On the other hand, these technological

Figure 2. A generic service model on DTV platform

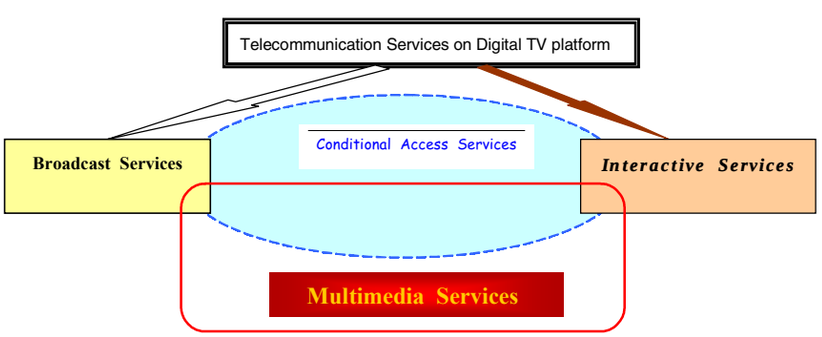
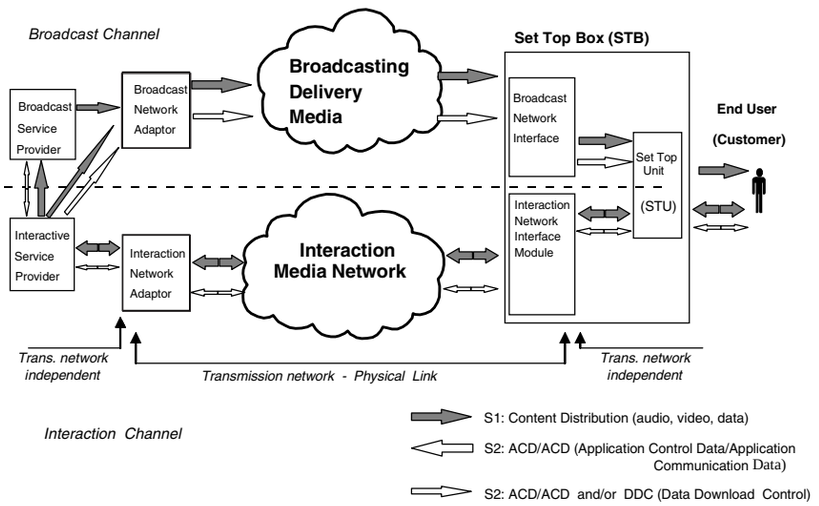


Figure 3. A generic model of DVB platform for interactive services



factors have a crucial impact on the network's ability to simultaneously manage services calling for different rates (Newman, 1996; Pagani, 2003). Moreover, interactivity requirements vary according to the respective service; that is, from the simple dispatch of a small amount of data to the network for ordering a programme (e.g., pay-per-view) to videophony, a service requiring symmetrical interactivity. The services will also need different type of links established between server and user and between subscribers themselves; for example, a specific point-to-point link between the server and the subscriber for a video on demand (VoD) service or flexible access to a large number of servers in transaction or information services, or the simple broadcast channel in broadcast TV service. It is thus possible to specify several categories of service, ranging from the most asymmetrical to the most symmetrical, as well as with local, unidirectional or bidirectional interactivity (Huffman, 2002).

Figure 4. The layer model of services on DTV platform

Interactive Services	Information	Interactive teletext Electronic Prog Guide Advertising Video yellow pages	Content Driven
	Participation	Direct response TV	
	Schedule	Pay per view Near video on demand Video on demand	
	Convenience	Tele-banking Real estate	Value Added
	Education	Lessons re-runs	
	Games	Trial games Network games	Enhanced TV
	Conversational	Video phone Video conference	
	Internet	Limited access Full access	

Besides, the services claim also for subscriber management functions, involving access conditioning, billing, means of payment, tiering of services, consumption statistics and so forth. These functions, in turn, will not only call for specific equipment but will also involve a big change in the profession of the operator, no matter what the network system (DVB-S, DVB-C or DVB-T).

Although the demand for new multimedia services is hard to evaluate and maybe there is no real killer application, the broadcasters and network operators in general tend to agree on an initial palette of services most likely to be offered. Based on the object and content of services aimed at a

residential audience, the palette of interactive multimedia services involves: leisure services, information and education services, and services for households (Fontaine, 1997). However, the technological convergence has impact on each traditional service-oriented sectors of the communications industry. Hence, as the entertainment, information, telecommunications and transaction sectors are becoming more dependent on each other, their products tend to be integrated. The resultant interactive multimedia services will constitute the layer model of telecommunication services to be provided on DTV platform (see Figure 4).

Taking into account a typology of telecommunication services to be provided via television medium and using a general prediction method for user demands (Hulicki,1998) one can estimate an early demand profile for a given subscriber type of a DTV platform.

Infrastructure for Provision of Services

It has been already mentioned that DVB seems to be one of the most important and insightful technologies for providing a personalized service environment. Its technical capabilities can be used to support the integration of DTV and the interactive multimedia services, and to meet future demands of users as well (Bancroft, 2001). The indispensable infrastructure for transporting DTV and multimedia services include both wire and wireless broadcast networks, and not only traditional TV network carriers (terrestrial, satellite and cable) are endeavouring to offer such services, but also new operators, who use for that purpose competitive technological options such as multichannel multipoint distribution service (MMDS) or microwave video distribution service (MVDS) (Dunne, 2000; Whitaker, 2003).

Looking at the infrastructure offering, it is important to note that the technical capacities of different network types comprising: satellite communications (TV Sat), cable (CATV) and terrestrial networks (diffusion TV), MMDS and public switched telecommunication networks (PSTN) might or might not be essentially different. Hence, the competitive position of each of these infrastructures in offering DTV, multimedia and interactive services can be apprehended by discussing the following aspects (de Bruin, 1999):

- the current position they hold in the residential market in terms of penetration rates
- the network deployment conditions; that is, the size of financial investments required for implementing suitable technology for providing these services and the time needed to build or modernise the networks
- the types of service they can or will be able to offer, conditioned by the optimal technical characteristics of the networks.

By exploring the existing environment that affects the development of digital television and related multimedia services and carrying out a comprehensive analysis of the networks in terms of service provision, network implementation and their potential technical evolution in the future, one should be able to answer the question: Which services on which networks?

Most market players agree that DTV will be distributed by all the traditional television carriers: terrestrial, satellite and cable networks, although an interesting alternative is new technological options; for example, wireless cable/MMDS (Scalise, 1999). Hence, the infrastructures for transporting DTV and/or multimedia services include both wire and wireless broadcast networks, and all telecoms operators are endeavouring to increase network capacity to be able to offer DTV and interactive multimedia services.

Currently, the major way for distributing DTV services is the broadcast of radio signals (see Figure 3). In some countries, however, the terrestrial network for broadcasting analog TV still dominates the television market, but the idea of using this infrastructure to transmit digital signals has received little encouragement, compared with cable and satellite alternatives (Hulicki, 2000). Unfortunately, digital terrestrial broadcasting comes up against congestion in the terrestrial spectrum and some other problems relating to control of the network (Spagat, 2002). On the other hand, a satellite TV broadcasting is a relatively simple arrangement. From the individual reception point of view, the advantages of this transmission mode include: simplicity, speed (immediate ensuring of large reception) and attractiveness; that is, relatively small cost of service implementation as well as a wide (international) coverage (Elbert, 1997). The market for collective reception, however, seems

fairly difficult for digital satellite services to break into (Huffman, 2002) in many parts of the world.

Another major way for distributing DTV services is supported by the wire networking infrastructure; that is, users can subscribe to two coexistent types of wire network: CATV networks and the telephone network. Originally, these networks performed clearly different functions: CATV networks are broadband and unidirectional, whereas telecommunications networks are switched, bi-directional and narrowband. Neither of these networks currently is capable of distributing complex interactive multimedia services: cable network operators lack the bi-directionality and even the switching (Thanos, 2001), while operators of telecommunications networks do not have the transmission band for transporting video services. In both cases, the local loop is in the forefront; on the one hand because it constitutes a technical bottleneck and, on the other, because it represents a major financial investment (de Bruin, 1999; Scalise, 1999).

The basic question concerning the development of MMDS networks is bound with the frequency zones to be used; that is, the number of channels and transmitter coverage vary with the allocation of frequencies, which is a highly regulated national concern (Hulicki, 2000). The digitization of an MMDS system calls for installation of specific equipment at the network head end and adaptation of the transmitter, but it is able to overcome the main drawback in this transmission mode; that is, limited channel capacity, and it may be perceived as a transition technology. In countries enjoying high cable network penetration, its development will undoubtedly remain limited, whereas in countries where CATV is encountering difficulties or where it is virtually unknown, the system is certainly of real interest (Furht, 1999).

Regardless of the transmission medium, the reception of DTV services calls for the installation of subscriber decoders (set top boxes, or STBs) for demodulating the digital signals to be displyed on a user's terminal screen (TV receiver), which itself will remain analog for the next few years. On the other hand, it seems to be obvious that future TV sets will access some computer services and a TV tuner will be incorporated in PCs. Nevertheless, computer and TV worlds are still deeply different. A TV screen is not suited for multimedia content and its text capabilities are very poor. The remote control does not satisfy the real user needs; that is, the interface tool is unable to make an easy navigation through programs, sites or multimedia content. Computing power and media storage are low, and it is hard to see how an interactive TV Guide could work in an easy, friendly manner with the embedded hardware of today.

Unlike a TV set, PC architecture is universal and cheap, but its life is short. A PC also has a smaller screen and is not yet adapted to high-speed multimedia. One can expect, however, that because of constantly adding or upgrading software and hardware by users, the capabilities of PCs will continue to grow. Hence, the convergence of PCs and TVs is underlying a large debate about future TV. Nevertheless, one can assume that PCs and TVs probably will be following parallel paths, but will not merge completely (Fontaine, 1997). Therefore, today, major TV companies have put the biggest investment on DVB and subscriber decoders. The role of the decoder is of foremost importance, as it represents the access to the end subscriber (Dreazen, 2002). The uniqueness or multiplicity of set top boxes, however, remains a key issue.

Creating and Delivering Services

Distribution of DTV and interactive multimedia services via satellite and/or terrestrial TV channels (e.g., MMDS or MVDS systems) truly seems to be a solution to fulfill the needs for broadband communication of the information age (Dunne, 2000). However, different requirements imposed by the various approaches to satellite communication systems have consequences on system design and its development. The trade-offs between maximum flexibility on one hand and complexity and cost on the other are always difficult to decide, since they will have an impact not only on the initial deployment of a system, but also on its future evolution and market acceptance (de Bruin, 1999). Moreover, the convergence of services and networks has changed requirements for forthcoming imaging formats (Boman, 2001; DAVIC, 1998).

As traditional networks begin to evolve towards new multiservice infrastructure, many different clients (TVs, home PCs and game consoles) will access content on the Internet. Because the traditional telecommunications market has been vertically integrated (de Bruin, 1999), with applications and services closely tied to the delivery channel, new solu-

tions, based on the horizontally layered concept that separates applications and services from the access and core networks, have to be developed. Besides, specific emphasis should be placed on the critical issues associated with a dual-band communication link concept; namely, a broad band on the forward path and a narrow or wide band on the return interactive path. In the future multimedia scenario, the integration of satellite resources with terrestrial networks will support the technical and economical feasibility of services via satellite and/or terrestrial TV. To develop multimedia services and products for different categories of users, several aspects have to be considered.

The process starts with service definition. In this phase, candidate applications have to be analyzed, targeting two main categories of users, namely "residential" and "business," from which the user profile could be derived. The next stage will include system design; that is, the typology of the overall system architecture for the operative system has to be assessed and designed. A specific effort should be placed on integrated distribution of services (DTV and multimedia together with interactive services) as well as on the service access scheme for the return channel operating in a narrow or wide band, aimed at identifying a powerful access protocol. In parallel, various alternatives of the return interactive channel can be considered and compared with the satellite solution. Then, a clear assessment of the system's economical viability will be possible. Different components of the system architecture should be analyzed in terms of cost competitiveness in the context of a wide and probable intensive expansion of services provided through an interactive DVB-like operative system. The objective of the analysis will not only be to define the suitability of such a technology choice but also to point out the applications and services that can be better exploited on the defined DVB system architecture.

In an attempt to cope with implementation aspects and design issues, service providers are faced with a dilemma. Not only must they choose an infrastructure that supports multiple services, but they must also select, from among a variety of last-mile access methods, how to deliver these multimedia services cost effectively now and in the future. Besides, when a new service succeeds, an initial deployment phase is usually followed by a sustained period of significant growth. The management systems must not only be able to cope with a high volume of initial network deployment activity, but also with the subsequent rapidly accelerating increase in the load (Dunne, 2000). Hence, in the service domain, one has to:

- analyze distribution of DTV programs bounded with delivery of advanced multimedia services to residential customers in a number of different areas: education (i.e., distance learning) and information (e.g., news on demand), entertainment (e.g., movie on demand, broadcast services) and commercial (e.g., home shopping)
- thoroughly evaluate the possibility of offering high-quality multimedia services with different levels of interactivity, ranging from no interaction (e.g., broadcast services) to a reasonably high level of interaction (i.e., distance learning and teleworking, also transaction services; e.g., teleshopping)
- analyze both the relationship between DVB and interactive services and the possibility of accessing interactive multimedia and Internet via different terminal equipment, from set-top boxes to PCs, taking careful consideration of the evolution of the former towards network computing devices.

In order to achieve the best overall system solution in the delivery platform domain, the following key issues should be addressed:

- optimization of both the service access on the return link in the narrow or wide band (in terms of protocols, transmission techniques, link budget trade-offs, etc.) and usage of downstream bandwidth for provision of interactive services bounded up with distribution of DTV programmes
- adoption of alternative solutions for the return channel in different access networks scenaria
- cost effectiveness of the user terminal (RF subsystem and set-top box) and viability of the adopted system choices for supporting new services
- effective integration of DVB platform on the surrounding interactive multimedia environment.

Apart from the overall system concept and its evolution, the introduction of such an integrated platform will have a large impact, even in the short term, due to the potential of DTV (multimedia market), and from the social one, due to the large number of actual and potential users of interactive services (e.g., the Internet). Besides, the introduction of multicast servers for multipoint applications should significantly increase the potential of the integrated DVB infrastructures, allowing interactive access to multimedia applications offered by content developers and service providers.

FUTURE TRENDS

In recent years, one can observe a convergence of various information and communication technologies on media market (Pagani, 2003). As a result of that process, the DTV sector is also subject to the convergence. Hence, the entertainment, information, telecommunications and transaction sectors of media market can play an important part in the development of new interactive multimedia services in the context of DTV. The market players from these traditional sectors are trying to develop activities beyond the scope of the core business and are competing to play the gatekeeper's role between sectors (Ghosh, 2002). At the same time, however, they also have to cooperate by launching joint ventures, in order to eliminate uncertainties in a return on investment, typical for new markets. Economies of scale can lead to cost reductions and, thus, to lower prices for customers. Moreover, combined investments can also lead to a general improvement of services. From the user's perspective, one of the positive and useful results of such integration on the basis of cooperation could be creation of a one-stop "shopping counter" through which all services from the various broadcasters could be provided. Such solution offers three important advantages; that is, the user does not have to sign up with every single service provider, there is no necessity to employ different modules to access the various services and, finally, competition will take place on the quality of service,rather than on the access to networks. From the broadcaster's point of view, the advantage of the open STB is that the network providers could still use its proprietary conditional access management system (CAMS). Hence, there is a number of questions and open problems that concern the provision of multimedia services on DTV platform; for example, the creation of an economic model of the market for DTV services, both existing and potential, that might be used for forecasting, a development of both the new electronic devices (STBs, integrated or DTV receivers) to be used at the customer premises and new interactive multimedia services (Newell, 2001). In general, because technological developments in the field of DTV have implications for the whole society, policy and decision makers in the government, industry and consumer organizations must assess these developments and influence them if necessary.

CONCLUSION

This contribution aimed to explore various aspects dealing with the provision of interactive and multimedia services on the DTV platform. Without pretending to be exhaustive, the article provides an overview of DVB technology and describes both the existing and potential multimedia services to be delivered on the DTV platform. It also examines the ability of the DTV infrastructure for provision of different services. Because this field is undergoing rapid development, underlying this contribution is also a question of the possible substitutions of services which previously were not substitutable. At the same time, an impact of the regulatory measures on the speed and success of the introduction of DTV and related multimedia services has been also discussed.

The scope of this article does not extend to offering conclusive answers to the above-mentioned questions or to resolving the outlined issues. In the meantime, the article is essentially a discussion document, providing a template for evaluating current state-of-the-art and conceptual frameworks that should be useful for addressing the questions to which the media market players must, in due course, resolve in order to remove barriers impeding progress towards a successful implementation of digital multimedia TV services.

REFERENCES

Bancroft, J. (2001). Fingerprinting: monitoring the use of media assets. *Proceedings of the International Broadcasting Convention Conference – IBC'01*, 55-63.

Boman, L. (2001). Ericsson's service network: A "melting pot" for creating and delivering mobile Internet service. *Ericsson Review*, 78, 62-67.

Burnett, R. et al. (Editors). (2004). *Perspectives on Multimedia: Communication, media and information techn*ology. New York: John Wiley & Sons.

Collins, G.W. (2001). *Fundamentals of digital television transmission*. New York: John Wiley & Sons.

DAVIC. (1998). Digital Audio-Visual Council 1.4 specification. Retrieved 2003 from *www.davic.org/*

de Bruin, R., & Smits, J. (1999). *Digital video broadcasting: technology, standards, and regulations*. Norwood: Artech House.

Dreazen, Y.J. (2002). FCC gives TV makers deadline of 2006 to roll out digital sets. *The Wall Street Journal*, Tues. Aug. 6, CCXL (26).

Dunne, E., & Sheppard, C. (2000). Network and service management in a broadband world. *Alcatel Telecommun. Rev.*, 4th Quarter, 262-268.

DVB/European Standards Institute. (1996). Support for use of scrambling and Conditional Access (CA) within digital broadcasting systems. *ETR 189*. Retrieved 2003 from *www.dvb.org/*

EBU. (1995). Functional model of a conditional access system. *EBU* (European Broadcast Union) *Technical Review*, winter 1995, 64-77.

Elbert, B. (1997). *The satellite communication applications handbook.* Norwood: Artech House.

ETR. (1996). *Digital Video Broadcasting (DVB);Guidelines for the use of the DVB Specification: Network independent protocols for interactive services (ETS 300 802)*.

ETSI. (2000). Digital Video Broadcasting (DVB); Interaction channel for satellite distribution systems. ETSI EN 301 790 V1.2.2 (2000-12).

Fontaine, G. et al. (1997°). *Internet and television*. Paris: IDATE Res. Rep., IDATE, September 1997.

Fontaine, G., & Hulicki, Z. (1997). Broadband infrastructures for digital television and multimedia services. *ACTS 97 - AC025 BIDS - Final Report*. IDATE, March 1997.

Furht, B., Westwater, R. & Ice, J. (1999). Multimedia broadcasting over the Internet: part II – video compression. *IEEE Multimedia*, Jan.-March, 85-89.

Ghosh, A.K. (2002). Addressing new security and privacy challenges. *IT Pro*, May-June, 10-11.

Huffman, F. (2002). Content distribution and delivery. *Tutorial, Proceedings of the 56th Annual NAB Broadcast Eng. Conference,* Las Vegas, NV.

Hulicki, Z. (1998). Modeling and dimensioning problems of the interaction channel for DVB systems. *Proceedings of the International Conference on Communication Technology*, S41–05–1-6.

Hulicki, Z. (2000). Digital TV platform – east European perspective. *Proceedings of the International Broadcasting Convention Conference – IBC'00*, 303-307.

Hulicki, Z. (2001). Integration of the interactive multimedia and DTV services for the purpose of distance learning. *Proceedings of the International Broadcasting Convention Conference – IBC'01*, 40-43.

Hulicki, Z. (2002). Security aspects in content delivery networks. *Proceedings of the 6th World Multiconference SCI'02 / ISAS'02*, 239-243.

Hulicki, Z., & Juszkiewicz, K. (1999). Internet on demand in DVB platform – performance modeling. *Proceedings of the 6th Polish Teletraffic Symposium*. 73-78.

Lotspiech, J., Nusser, S., & Pestoni, F. (2002). Broadcast encryption's bright future. *IEEE Computer, 35*(8), August, 57-63.

Mauthe, A., & Thomas P. (2004). *Professional content management systems: Handling digital media assets*. New York: John Wiley & Sons.

Newell, J. (2001). The DVB MHP Internet access profile. *Proceedings of the International Broadcasting Convention Conference – IBC'01,* 266-271.

Newman, W.M., & Lamming, M.G. (1996). *Interactive system design*. New York: Addison-Wesley.

Pagani, M. (2003). *Multimedia and interactive digital TV: Managing the opportunities created by digital convergence.* Hershey: Idea Publishing Group.

Raghavan, S.V., & Tripathi, S.K. (1998). *Networked multimedia systems: concepts, architecture, and design.* Upper Saddle River: Prentice Hall.

Rodriguez, A., & Mitaru, A. (2001). File security and rights management in a network content server system. *Proceedings of the International Broadcasting Convention Conference – IBC'01,* 78-82.

Scalise, F., Gill, D., & Faria, G., et al. (1999). Wireless terrestrial interactive: a new TV system based on DVB-T and SFDMA, proposed and demonstrated by iTTi project. *Proceedings of the International Broadcasting Convention Conference – IBC'99,* 26-33.

Seffah, A., & Javahery, H. (Eds.). (2004) *Multiple user interfaces: Cross-platform spplications and context-aware interfaces*. New York: John Wiley & Sons.

Spagat, E. (2002). The revival of DTV. *The Wall Street Journal*, August 1, CCXL (23).

Tatipamula, M., & Khasnabish, B. (Editors). (1998). *Multimedia communications networks. Technologies and services*. Norwood: Artech House.

Thanos, D., & Konstantas, D. (2001). A model for the commercial dissemination of video over open networks. *Proceedings of the International Broadcasting Convention Conference – IBC'01,* 83-94.

Wang, Z., & Crowcroft, J. (1996). Quality-of-Service routing for supporting multimedia applications. *IEEE Journal on Selected Areas in Communications*, *14*(7), 1228-1234.

Whitaker, J., & Benson, B. (2003). *Standard handbook of video and television engineering*. New York: McGraw-Hill.

KEY TERMS

Broadcast TV Services: Television services that provide a continuous flow of information distributed from a central source to a large number of users.

Conditional Access (CA) Services: Television services that allow only authorized users to select, receive, decrypt and watch a particular programming package.

Content-Driven Services: Television services to be provided depending on the content.

Digital Video Broadcasting (DVB): The European standard for the development of DTV.

DTV: Broadcasting of television signals by means of digital techniques, used for the provision of TV services.

Enhanced TV: A television that provides subscribers with the means for bi-directional communication with real-time, end-to-end information transfer.

Interactive Services: Telecommunication services that provide users with the ability to control and influence the subjects of communication.

Multimedia Communication: A new, advanced way of communication that allows any of the traditional information forms (including their integration) to be employed in the communication process.

Set Top Box (STB): A decoder for demodulating the digital signals to be displyed on a TV receiver screen.

Value-Added Services: Telecommunication services with the routing capability and the established additional functionality.

Multimedia Content Representation Technologies

Ali R. Hurson
The Pennsylvania State University, USA

Bo Yang
The Pennsylvania State University, USA

INTRODUCTION

Multimedia: Promises and Challenges

In recent years, the rapid expansion of multimedia applications, partly due to the exponential growth of the Internet, has proliferated over the daily life of Internet users. Consequently, research on multimedia technologies is of increasing importance in computer society. In contrast with traditional text-based systems, multimedia applications usually incorporate much more powerful descriptions of human thought – video, audio and images (Auffret, Foote, Li & Shahraray, 1999). Moreover, the large collections of data in multimedia systems make it possible to resolve more complex data operations, such as imprecise query or content-based retrieval. For instance, image database systems may accept an example picture and return the most similar images of the example (Cox, Miller & Minka, 2000, Huang, Chang & Huang, 2003). However, the conveniences of multimedia applications come at the expense of new challenges to the existing data management schemes:

- Multimedia applications generally require more resources; however, the storage space and processing power are limited in many practical systems; for example, mobile devices and wireless networks (Lim & Hurson, 2002). Due to the large size of multimedia databases and complicated operations of multimedia applications, new methods are needed to facilitate efficient accessing and processing of multimedia data while considering the technological constraints (Bourgeois, Mory & Spies, 2003).
- There is a gap between user perception and physical representation of multimedia data. Users often browse and desire to access multimedia data at the object level ("entities" such as human beings, animals or buildings). However, the existing multimedia-retrieval systems tend to represent multimedia data based on their lower-level features ("characteristics" such as color patterns and textures), with less emphases on combining these features into objects (Hsu, Chua & Pung, 2000). This representation gap often leads to unexpected retrieval results. The representation of multimedia data according to a human's perspective is one of the focuses in recent research activities; however, no existing systems provide automated identification or classification of objects from general multimedia collections (Kim & Kim, 2002).
- The collections of multimedia data are often diverse and poorly indexed (Huang et al., 2002). In a distributed environment, due to the autonomy and heterogeneity of data sources, multimedia objects are often represented in heterogeneous formats (Kwon, Choi, Bisdikian & Naghshineh, 2003). The difference in data formats further leads to the difficulty of incorporating multimedia objects within a unique indexing framework (Auffret et al., 1999).
- Last but not least, present research on content-based multimedia retrieval is based on features. These features are extracted from the audio/video streams or image pixels, with the empirical or heuristic selection, and then combined into vectors according to the application criteria (Hershey & Movellan, 1999). Due to

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the application-specific multimedia formats, this paradigm of multimedia data management lacks scalability, accuracy, efficiency and robustness (Westermann & Klas, 2003).

Representation: The Foundation of Multimedia Data Management

Successful storage and access of multimedia data, especially in a distributed heterogeneous database environment, require careful analysis of the following issues:

- Efficient representation of multimedia entities in databases
- Proper indexing architecture for the multimedia databases
- Proper and efficient technique to browse and/or query objects in multimedia database systems.

Among these three issues, multimedia representation provides the foundation for indexing, classification and query processing. The suitable representation of multimedia entities has significant impact on the efficiency of multimedia indexing and retrieval (Huang et al., 2003). For instance, object-level representation usually provides more convenient content-based indexing on multimedia data than pixel-level representation (Kim & Kim, 2002). Similarly, queries are resolved within the representation domains of multimedia data, either at the object level or pixel level (Hsu et al., 2000). The nearest-neighbor searching schemes are usually based on careful analysis of multimedia representation – the knowledge of data contents and organization in multimedia systems (Yu & Zhang, 2000; Li et al., 2003).

The remaining part of this article is organized into three sections: First, we offer the background and related work. Then, we introduce the concepts of semantic-based multimedia representation approach and compare it with the existing non-semantic-based approaches. Finally, we discuss the future trends in multimedia representation and draw this article into a conclusion.

BACKGROUND

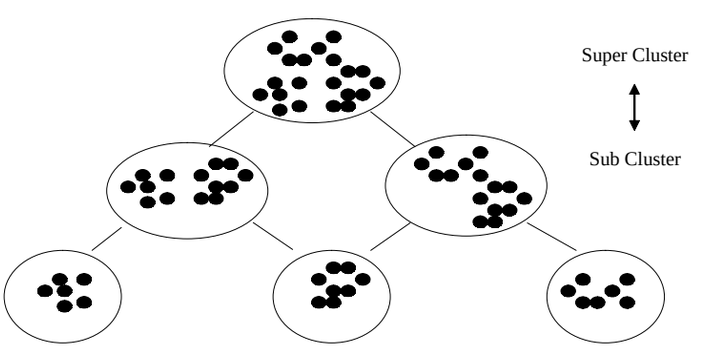
Preliminaries of Multimedia Representation

The main goal of multimedia representation is to obtain a concise content description during the analysis of multimedia objects. Representation approaches as advanced in the literature are classified into four groups: clustering-based, representative-region-based, decision-tree-based and annotation-based.

Clustering-Based Approach

The clustering-based approach recursively merges content-similar multimedia objects into clusters with human intervention or automated classification algorithms while obtaining the representation of these multimedia objects. There are two types of clustering schemes: *supervised* and *unsupervised* (Kim & Kim,

Figure 1. The decomposition of clusters



2002). The supervised clustering scheme utilizes the user's knowledge and input to cluster multimedia objects, so it is not a general-purpose approach. As expected, the unsupervised clustering scheme does not need interaction with the user. Hence, it is an ideal way to cluster unknown multimedia data automatically (Heisele & Ritter, 1999). Here we only discuss the unsupervised clustering scheme, because of its advantages.

In the clustering-based approach, the cluster of a multimedia object indicates its content (Rezaee, Zwet & Lelieveldt, 2000). The clusters are organized in a hierarchical fashion – a super cluster may be decomposed into several sub clusters and represented as the union of sub clusters (Figure 1). New characteristics are employed in the decomposition process to indicate the differences between sub clusters. Consequently, a sub cluster inherits the characteristics from its super cluster while maintaining its individual contents (Huang et al., 2003).

Representative-Region-Based Approach

The representative-region-based approach selects several representative regions from a multimedia object and constructs a simple description of this object based on the selected regions. The representative regions are some small areas with the most notable characteristics of the whole object. In case of an image, the representative regions can be areas that the color changes markedly, or areas that the texture varies greatly and so forth.

The representative-region-based approach is performed as a sequence of three steps:

- Region selection: The original multimedia object consists of many small regions. Hence, the selection of representative regions is the process of analyzing the changes in those small regions. The difference with the neighboring regions is quantified as a numerical value to represent a region. Finally, based on such a quantitative value, the regions are ordered, and the most notable regions are selected.
- Function application: The foundation of the function application process is the Expectation Maximization (EM) algorithm (Ko,& Byun, 2002). The EM algorithm is used to find the maximum likelihood function estimates when the multimedia object is represented by a small number of selected regions. The EM algorithm is divided into two steps: E-step and M-step. In the E-step, the features for the unselected regions are estimated. In the M-step, the system computes the maximum-likelihood function estimates using the features obtained in the E-step. The two steps alternate until the functions are close enough to the original features in the unselected regions.
- Content representation: The content representation is the process that integrates the selected regions into a simple description that represents the content of the multimedia object. It should be noted that the simple description is not necessarily an exhaustive representation of the content. However, as reported in the literature, the overall accuracy of expressing multimedia contents is acceptable (Jing, Li, Zhang & Zhang, 2002).

Decision-Tree-Based Approach

The decision-tree-based approach is the process of obtaining content of multimedia objects through decision rules (MacArthur, Brodley & Shyu, 2000). The decision rules are automatically generated standards that indicate the relationship between multimedia features and content information. In the process of comparing the multimedia objects with decision rules, some tree structures – decision trees – are constructed (Simard, Saatchi & DeGrandi, 2000).

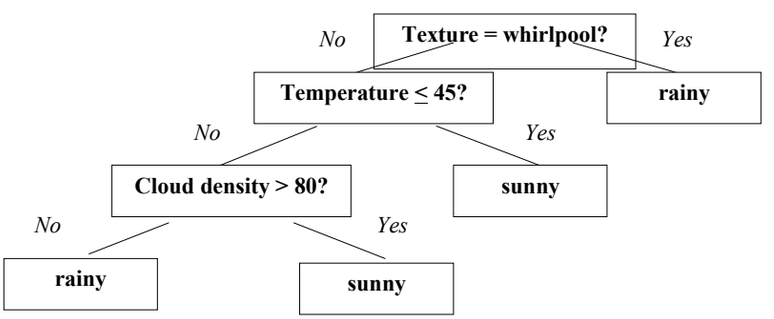
The decision-tree-based approach is mostly applicable in application domains where decision rules can be used as standard facts to classify the multimedia objects (Park, 1999). For example, in a weather forecasting application, the satellite-cloud images are categorized as rainy and cloudy according to features such as cloud density and texture. Different combinations of feature values are related to different weathers (Table 1). A series of decision rules are derived to indicate these relationships (Figure 2). And the final conclusions are the contents of the multimedia objects.

The decision-tree-based approach can improve its accuracy and precision as the number of analyzed multimedia objects increases (Jeong & Nedevschi, 2003). Since the decision rules are obtained from statistical analysis of multimedia

Table 1. The features of cloud images

Temperature	Cloud density	Texture	Weather
45	90	plain	rainy
50	60	whirlpool	rainy
65	75	plain	sunny
38	80	plain	rainy
77	50	plain	sunny
53	85	plain	rainy
67	100	whirlpool	rainy

Figure 2. A decision tree for predicting weathers from cloud images



objects, more sample objects will result in improved accuracy (MacArthur et al., 2000).

Annotation-Based Approach

Annotation is the descriptive text attached to multimedia objects. Traditional multimedia database systems employ manual annotations to facilitate content-based retrieval (Benitez, 2002). Due to the explosive expansion of multimedia applications, it is both time-consuming and impractical to obtain accurate manual annotations for every multimedia object (Auffret et al., 1999). Hence, automated multimedia annotation is becoming a hotspot in recent research literature. However, even though humans can easily recognize the contents of multimedia data through browsing, building an automated system that generates annotations is very challenging. In a distributed heterogeneous environment, the heterogeneity of local databases introduces additional complexity to the goal of obtaining accurate annotations (Li et al., 2003).

Semantic analysis can be employed in annotation-based approach to obtain extended content description from multimedia annotations. For instance, an image containing "flowers" and "smiling faces" may be properly annotated as "happiness." In addition, a more complex concept may be deduced from the combination of several simpler annotations. For example, the combination of "boys," "playground" and "soccer" may express the concept "football game."

Comparison of Representation Approaches

The different rationales of these multimedia-representation approaches lead to their strengths and weaknesses in different application domains. Here these approaches are compared under the consideration of various performance merits (Table 2).

The approaches do not consider the semantic contents that may exist in the multimedia objects. Hence, they are collectively called "non-semantic-

Table 2. Comparison of representation approaches

Performance Merit	*Clustering*	*Representative Region*	*Decision Tree*	*Annotation*
Rationale	Searching pixel-by-pixel, recognizing all details	Selecting representative regions	Treating annotations as multimedia contents	Using annotations as standard facts
Reliability & Accuracy	Reliable and accurate	Lack of robustness	Depending on the accuracy of annotations	Robust and self-learning
Time Complexity	Exhaustive, very time consuming	Most time is spent on region selection	Fast text processing	Time is spent on decision rules and feedback
Space Complexity	Large space requirement	Relatively small space requirement	Very small storage needed	Only need storage for decision rules
Application Domain	Suitable for all application domains	The objects that can be represented by regions	Need annotations as basis	Restricted to certain applications
Implementation Complexity	Easy to classify objects into clusters	Difficult to choose proper regions	Easily obtaining content from annotations	Difficult to obtain proper decision rules

based" approaches. Due to the lack of semantic analysis, they usually have the following limitations:

- Ambiguity: The multimedia contents are represented as numbers that are not easily understood or modified.
- Lack of robustness and scalability: Each approach is suitable for some specific application domains, and achieves the best performance only when particular data formats are considered. None of them has the capability of accommodating multimedia data of any format from heterogeneous data sources.

MAIN FOCUS OF THE ARTICLE

The limitations of non-semantic-based approaches lead to the research on semantic-based multimedia-representation methods. One of the promising models in the literature is the summary-schemas model (SSM).

Summary Schemas Model

The SSM is a content-aware organization prototype that enables imprecise queries on distributed heterogeneous data sources (Ngamsuriyaroj, 2002). It provides a scalable content-aware indexing method based on the hierarchy of summary schemas, which comprises three major components: a thesaurus, a collection of autonomous local nodes and a set of summary-schemas nodes (Figure 3).

The thesaurus provides an automatic taxonomy that categorizes the standard accessing terms and defines their semantic relationships. A local node is a physical database containing the multimedia data. With the help of the thesaurus, the data items in local databases are classified into proper categories and represented with abstract and semantically equivalent summaries. A summary-schemas node is a virtual entity concisely describing the semantic contents of its child/children node(s). More detailed descriptions can be found in Jiao and Hurson (2004).

Figure 3. Summary Schemas Model

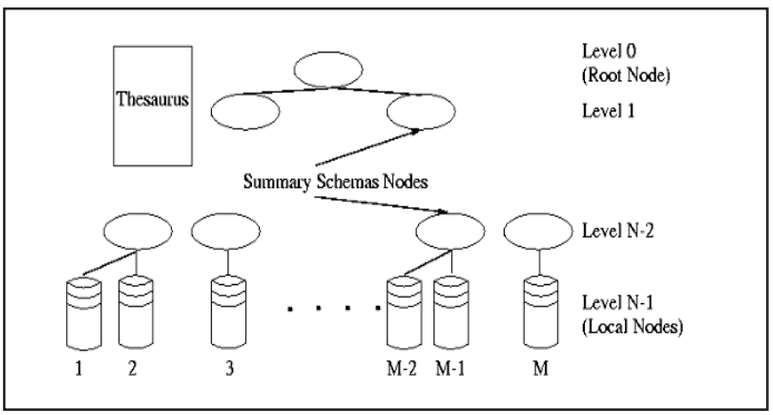
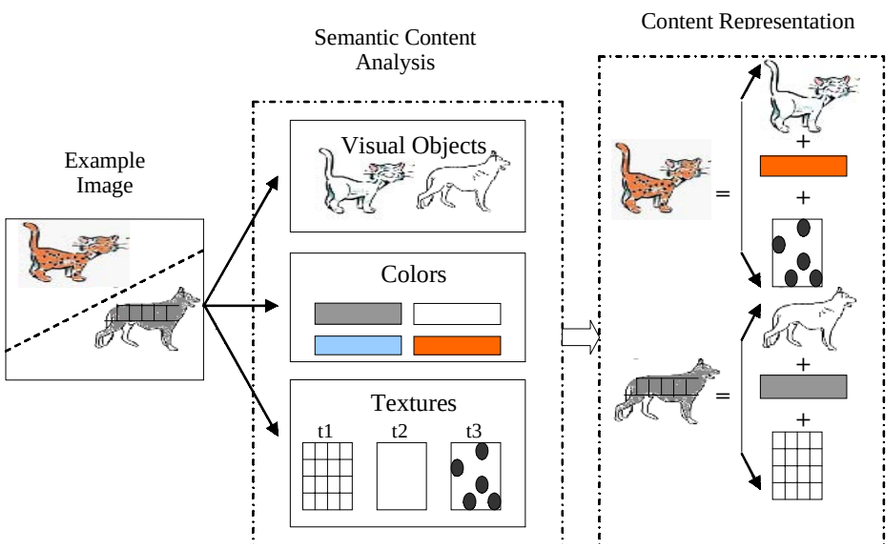


Figure 4. Semantic content components of image objects

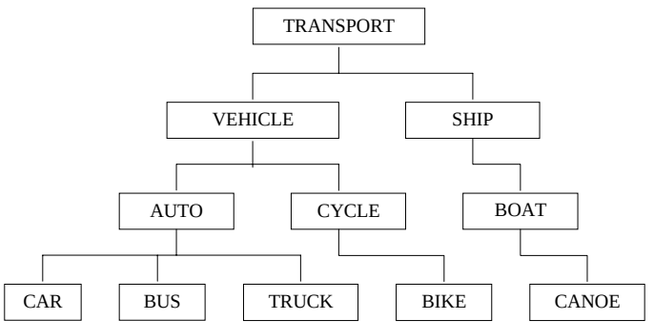


To represent the contents of multimedia objects in a computer-friendly structural fashion, the SSM organizes multimedia objects into layers according to their semantic contents. A multimedia object – say, an image – can be considered as the combination of a set of elementary entities, such as animals, vehicles and buildings. And each elementary entity can be described using some logic predicates that indicate the mapping of the elementary entity on different features. For instance, the visual elementary objects in Figure 4 are dog and cat. The possible color is grey, white, blue or brown. The texture pattern is $texture_1$, $texture_2$ or $texture_3$. Hence, the example image in Figure 4 can be represented as the combination of visual objects, colors and textures, such as (cat $\wedge$ brown $\wedge$ t3) $\vee$ (dog $\wedge$ grey $\wedge$ t1).

Non-Semantic-Based Methods vs. Semantic-Based Scheme

In contrast with the multimedia-representation approaches mentioned earlier, the SSM employs a unique semantic-based scheme to facilitate multimedia representation and organization. A multime-

Figure 5. The SSM hierarchy for multimedia objects



dia object is considered as a combination of logic terms that represents its semantic content. The analysis of multimedia contents is then converted to the evaluation of logic terms and their combinations. This content-representation approach has the following advantages:

- The semantic-based descriptions provide a convenient way of representing multimedia contents precisely and concisely. Easy and consistent representation of the elementary objects based on their semantic features simplifies the content representation of complex objects using logic computations – the logic representation of multimedia contents is often more concise than feature vector, which is widely used in non-semantic-based approaches.
- Compared with non-semantic-based representation, the semantic-based scheme integrates multimedia data of various formats into a unified logical format. This also allows the SSM to organize multimedia objects regardless of their representation (data formats such as MPEG) uniformly, according to their contents. In addition, different media types (video, audio, image and text) can be integrated under the SSM umbrella, regardless of their physical differences.
- The semantic-based logic representation provides a mathematical foundation for operations such as similarity comparison and optimization. Based on the equivalence of logic terms, the semantically similar objects can be easily found and grouped into same clusters to facilitate data retrieval. In addition, mathematical techniques can be used to optimize the semantic-based logic representation of multimedia entities – this by default could result in better performance and space utilization.
- The semantic-based representation scheme allows one to organize multimedia objects in a hierarchical fashion based on the SSM infrastructure (Figure 5). The lowest level of the SSM hierarchy comprises multimedia objects, while the higher levels consist of summary schemas that abstractly describe the semantic contents of multimedia objects. Due to the descriptive capability of summary schemas, this semantic-based method normally achieves more representation accuracy than non-semantic-based approaches.

FUTURE TRENDS AND CONCLUSION

The literature has reported considerable research on multimedia technologies. One of the fundamental research areas is the content representation of multimedia objects. Various non-semantic-based multimedia-representation approaches have been proposed in the literature, such as clustering-based approach, representative-region-based approach, decision-tree-based approach and annotation-based approach. Recent research results also show some burgeoning trends in multimedia-content representation:

- Multimedia-content processing through cross-modal association (Westermann,& Klas, 2003; Li et al., 2003).
- Content representation under the consideration of security (Adelsbach et al., 2003; Lin & Chang, 2001).

- Wireless environment and its impact on multimedia representation (Bourgeois et al., 2003; Kwon et al., 2003).

This article briefly overviewed the concepts of multimedia representation and introduced a novel semantic-based representation scheme – SSM. As multimedia applications keep proliferating through the Internet, the research on content representation will become more and more important.

REFERENCES

Adelsbach, A., Katzenbeisser, S., & Veith, H. (2003). Watermarking schemes provably secure against copy and ambiguity attacks. *ACM Workshop on Digital Rights Management,* 111-119.

Auffret, G., Foote, J., Li, & Shahraray C. (1999). Multimedia access and retrieval (panel session): The state of the art and future directions. *ACM Multimedia, 1*, 443-445.

Benitez, A.B. (2002). Semantic knowledge construction from annotated image collections. *IEEE Conference on Multimedia and Expo, 2*, 205-208.

Bourgeois, J., Mory, E., & Spies, F. (2003). Video transmission adaptation on mobile devices. *Journal of Systems Architecture, 49*(1), 475-484.

Cox, I.J., Miller, M.L., & Minka, T.P. (2000). The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing, 9*(1), 20-37.

Heisele, B., & Ritter, W. (1999). Segmentation of range and intensity image sequences by clustering. *International Conference on Information Intelligence and Systems*, 223-225.

Hershey, J., & Movellan, J. (1999). Using audio-visual synchrony to locate sounds. *Advances in Neural Information Processing Systems*, 813-819.

Hsu, W., Chua, T.S., & Pung, H.K. (2000). Approximating content-based object-level image retrieval. *Multimedia Tools and Applications, 12*(1), 59-79.

Huang, Y., Chang, T., & Huang, C. (2003). A fuzzy feature clustering with relevance feedback approach to content-based image retrieval. *IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems,* 57-62.

Jeong, P., & Nedevschi, S. (2003). Intelligent road detection based on local averaging classifier in real-time environments. *International Conference on Image Analysis and Processing,* 245-249.

Jiao, Y. & Hurson, A.R. (2004). Application of mobile agents in mobile data access systems – A prototype. *Journal of Database Management, 15*(4), 2004.

Jing, F., Li, M., Zhang, H., & Zhang, B. (2002). Region-based relevance feedback in image retrieval. *IEEE Symposium on Circuits and Systems,* 26-29.

Kim, J.B., & Kim, H.J. (2002). Unsupervised moving object segmentation and recognition using clustering and a neural network. *International Conference on Neural Networks, 2*, 1240-1245.

Ko B., & Byun, H. (2002). Integrated region-based retrieval using region's spatial relationships. *International Conference on Pattern Recognition,* 196-199.

Kwon, T., Choi, Y., Bisdikian, C., & Naghshineh, M. (2003). Qos provisioning in wireless/mobile multimedia networks using an adaptive framework. *Wireless Networks,* 51-59.

Li, B., Goh, K., & Chang, E.Y. (2003). Confidence-based dynamic ensemble for image annotation and semantics discovery. *ACM Multimedia,* 195-206.

Li, D., Dimitrova, N., Li, M., & Sethi, I.K. (2003). Multimedia content processing through cross-modal association. *ACM Multimedia,* 604-611.

Lim, J.B., & Hurson, A.R. (2002). Transaction processing in mobile, heterogeneous database systems. *IEEE Transaction on Knowledge and Data Engineering, 14*(6), 1330-1346.

Lin, C., & Chang, S. (2001). SARI: Self-authentication-and-recovery image watermarking system. *ACM Multimedia,* 628-629.

MacArthur, S.D., Brodley, C.E., & Shyu, C. (2000). Relevance feedback decision trees in content-based image retrieval. *IEEE Workshop on Content-based Access of Image and Video Libraries,* 68-72.

Ngamsuriyaroj, S., Hurson, A.R., & Keefe, T.F. (2002). Authorization model for summary schemas model. *International Database Engineering and Applications Symposium,* 182-191.

Park, I.K. (1999). Perceptual grouping of 3D features in aerial image using decision tree classifier. *International Conference on Image Processing, 1,* 31-35.

Rezaee, M.R., Zwet, P.M., & Lelieveldt, B.P. (2000). A multiresolution image segmentation technique based on pyramidal segmentation and fuzzy clustering. *IEEE Transactions on Image Processing, 9*(7), 1238-248.

Simard, M., Saatchi, S.S., & DeGrandi, G. (2000). The use of decision tree and multiscale texture for classification of JERS-1 SAR data over tropical forest. *IEEE Transactions on Geoscience and Remote Sensing, 38*(5), 2310–2321.

Westermann, U., & Klas, W. (2003). An analysis of XML database solutions for management of MPEG-7 media descriptions. *ACM Computing Surveys,* 331-373.

Yu, D., & Zhang, A. (2000). Clustertree: Integration of cluster representation and nearest neighbor search for image databases. *IEEE Conference on Multimedia and Expo, 3,* 1713-1716.

KEY TERMS

Annotation: Descriptive text attached to multimedia objects.

Cluster: A group of content-similar multimedia objects.

Decision Rule: Automatically generated standards that indicate the relationship between multimedia features and content information.

Elementary Entity: Data entities that semantically represent basic objects.

Representative Region: Areas with the most notable characteristics of a multimedia object.

Semantic-Based Representation: Describing multimedia content using semantic terms.

Summary-Schemas Model: A content-aware organization prototype that enables imprecise queries on distributed heterogeneous data sources.

Multimedia Data Mining Concept

Janusz Swierzowicz
Rzeszow University of Technology, Poland

INTRODUCTION

The development of information technology is particularly noticeable in the methods and techniques of data acquisition, high-performance computing, and bandwidth frequency. According to a newly observed phenomenon, called a storage low (Fayyad & Uthurusamy, 2002), the capacity of digital data storage is doubled every 9 months with respect to the price. Data can be stored in many forms of digital media, for example, still images taken by a digital camera, MP3 songs, or MPEG videos from desktops, cell phones, or video cameras. Such data exceeds the total cumulative handwriting and printing during all of recorded human history (Fayyad, 2001). According to current analysis carried out by IBM Almaden Research (Swierzowicz, 2002), data volumes are growing at different speeds. The fastest one is Internet-resource growth: It will achieve the digital online threshold of exabytes within a few years (Liautaud, 2001). In these fast-growing volumes of data environments, restrictions are connected with a human's low data-complexity and dimensionality analysis. Investigations on combining different media data, multimedia**,** into one application have begun as early as the 1960s, when text and images were combined in a document. During the research and development process, audio, video, and animation were synchronized using a time line to specify when they should be played (Rowe & Jain, 2004). Since the middle 1990s, the problems of multimedia data capture, storage, transmission, and presentation have extensively been investigated. Over the past few years, research on multimedia standards (e.g., MPEG-4, X3D, MPEG-7) has continued to grow. These standards are adapted to represent very complex multimedia data sets; can transparently handle sound, images, videos, and 3-D (three-dimensional) objects combined with events, synchronization, and scripting languages; and can describe the content of any multimedia object. Different algorithms need to be used in multimedia distribution and multimedia database applications. An example is an image database that stores pictures of birds and a sound database that stores recordings of birds (Kossmann, 2000). The distributed query that asks for "top ten different kinds of birds that have black feathers and a high voice" is described there by Kossmann (2000, p.436).

One of the results of the inexorable growth of multimedia data volumes and complexity is a data overload problem. It is impossible to solve the data overload issue in a human manner; it takes strong effort to use intelligent and automatic software tools for turning rough data into valuable information and information into knowledge.

Data mining is one of the central activities associated with understanding, navigating, and exploiting the world of digital data. It is an intelligent and automatic process of identifying and discovering useful structures in data such as patterns, models, and relations. We can consider data mining as a part of the overall knowledge discovery in data processes. Kantardzic (2003, p.5) defines data mining as "a process of discovering various models, summaries, and derived values from a given collection of data.". It should be an iterative and carefully planned process of using proper analytic techniques to extract hidden, valuable information.

The article begins with a short introduction to data mining, considering different kinds of data, both structured as well as semistructured and unstructured. It emphasizes the special role of multimedia data mining. Then, it presents a short overview of goals, methods, and techniques used in multimedia data mining. This section focuses on a brief discussion on supervised and unsupervised classification, uncovering interesting rules, decision trees, artificial neural networks, and rough-neural computing. The next section presents advantages offered by multimedia data mining and examples of practical and successful applications. It also contains a list of application domains. The following section describes multimedia data-mining critical issues and summa-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

rizes main multimedia data-mining advantages and disadvantages.

NEED FOR MULTIMEDIA DATA MINING

Data mining is essential as we struggle to solve data overload and complexity issues. With the fastest acceleration of off-line data resources on the Internet, the WWW (World Wide Web) is a natural area for using data-mining techniques to automatically discover and extract actionable information from Web documents and services. These techniques are named Web mining. We also consider text mining as a data-mining task that helps us summarize, cluster, classify, and find similar text documents in a set of documents. Due to advances in informational technology and high-performance computing, very large sets of images such as digital or digitalized photographs, medical images, satellite images, digital sky surveys, images from computer simulations, and images generated in many scientific disciplines are becoming available. The method that deals with the extraction of implicit knowledge, image data relationships, and other patterns not explicitly stored in the image databases is called image mining (Zhang, Hsu, & Li Lee, 2001a). A main issue of image mining is dealing with relative data, implicit spatial information, and multiple interpretations of the same visual patterns. We can consider the application-oriented functional approach and the image-driven approach. In the latter, one the following hierarchical layers are established (Zhang, Hsu, & Li Lee, 2001b): the lower layer that consists of pixel and object information, and the higher layer that takes into consideration domain knowledge to generate semantic concepts from the lower layer and incorporates them with related alphanumerical data to discover domain knowledge.

The main aim of the multimedia data mining is to extract interesting knowledge and understand semantics captured in multimedia data that contain correlated images, audio, video, and text.

Multimedia databases, containing combinations of various data types, could be first integrated via distributed multimedia processors and then mined, or one could apply data-mining tools on the homogenous databases and then combine the results of the various data miners (Thuraisingham, 2002).

GOALS, METHODS, AND TECHNIQUES USED IN MULTIMEDIA DATA MINING

One of the most popular goals in data mining is ordering or dissecting a set of objects described by high-dimensional data into small comprehensive units, classes, substructures, or parts. These substructures give better understanding and control, and can assign a new situation to one of these classes based on suitable information, which can be classified as supervised or unsupervised. In the former classification, each object originates from one of the predefined classes and is described by a data vector (Bock, 2002). But it is unknown to which class the object belongs, and this class must be reconstructed from the data vector. In unsupervised classification (clustering), a new object is classified into a cluster of objects according to the object content without a priori knowledge. It is often used in the early stages of the multimedia data-mining processes.

If a goal of multimedia data mining can be expressed as uncovering interesting rules, an association-rule method is used. An association rule takes a form of an implication $X \Rightarrow Y$, where X denotes antecedent of the rule, Y denotes the consequent of the rule, X and Y belong to the set of objects (item set) I, $X \cap Y = \Phi$, and D denotes a set of cases (Zhang et al., 2001a). We can determine two parameters named support, s, and confidence, c. The rule $X \Rightarrow Y$ has support s in D, where s% of the data cases in D contains both X and Y, and the rule holds confidence c in D, where c% of the data cases in D that support X also support Y. Association-rule mining selects rules that have support greater than some user-specified minimum support threshold (typically around 10^{-2} to 10^{-4}), and the confidence of the rule is at least a given (from 0 to 1) confidence threshold (Mannila, 2002). A typical association-rule mining algorithm works in two steps. The first step finds all large item sets that meet the minimum support constraint. The second step generates rules from all large item sets that satisfy the minimum confidence constraints.

A natural structure of knowledge is a decision tree. Each node in such a tree is associated with a

test on the values on an attribute, each edge from a node is labeled with a particular value of the attribute, and each leaf of the tree is associated with a value of the class (Quinlan, 1986). However, when the values of attributes for the description change slightly, the decision associated with the previous description can vary greatly. It is a reason to introduce fuzziness in decision trees to obtain fuzzy decision trees (Marsala, 2000). A fuzzy decision-tree method, equivalent to a set of fuzzy rules "if…then," represents natural and understandable knowledge (Detyniecki & Marsala , 2002).

In case the goal of multimedia data mining is pattern recognition or trend prediction with limited domain knowledge, the artificial neural-network approach can be applied to construct a model of the data. Artificial neural networks can be viewed as highly distributed, parallel computing systems consisting of a large number of simple processors (similar to neurons) with many weighted interconnections. "Neural network models attempt to use some organizational principles (such as learning, generalization, adaptivity, fault tolerance, distributed representation, and computation) in a network of weighted, directed graphs in which the nodes are artificial neurons, and directed edges (with weights) are connections between neuron outputs and neuron inputs" (Jain, Duin & Mao, 2000, p.9). These networks have the ability to learn complex, nonlinear input-output relationships, and use sequential training procedures.

For the need of dimensionality reduction, principal-component analysis (PCA) often is performed (Herena, Paquet, & le Roux, 2003; Skowron & Swinarski, 2004). In this method, the square covariance matrix, which characterizes the training data set, is computed. Next, the eigenvalues are evaluated and arranged in decreasing order with corresponding eigenvectors. Then the optimal linear transformation is provided to transform the n-dimensional space into m-dimensional space, where $m \leq n$, and m is the number of the most dominant, principal eigenvalues, that corresponds to the importance of each dimension. Dimensions corresponding to the smallest eigenvalues are neglected. The optimal transformation matrix minimizes the mean, last square-reconstruction error. In addition to PCA, rough-set theory can be applied for choosing eligible principal components, which describe all concepts in a data set, for classification. An appropriate algorithm of feature extraction and selection using PCA and rough sets is presented by Skowron & Swinarski.

ADVANTAGES OFFERED BY MULTIMEDIA DATA MINING

In multimedia data mining, classification is mainly interpreted as object recognition. Object models (e.g., letters or digits) are known a priori, and an automatic recognition system finds letters or digits from handwritten or scanned documents. Other examples are the identification of images or scenarios on the basis of sets of visual data from photos, satellites, or aero observations; the finding of common patterns in a set of images; and the identification of speakers and words in speech recognition. Image association-rule mining is used for finding associations between structures and functions of the human brain. One of the most promising applications of multimedia data mining is biometrics, which refers to the automatic identification of an individual by using certain physiological or behavioural traits associated with the person (Jain & Ross, 2004). It combines many human traits of the hand (hand geometry, fingerprints, or palm prints), eye (iris, retina), face (image or facial thermogram), ear, voice, gait, and signature to identify of an unknown user or verify a claimed identity. Biometrical systems must solve numerous problems of noise in biometric data, the modification of sensor characteristics, spoof, and replay: attacks in various real-life applications. A major area of research within biometric signal processing is face recognition. A face-detection system works with the edge features of greyscale, still images and the modified Hausdorff distance as described by Jesorsky, Kirchberg, and Frischolz (2001). It is used as a similarity measure between a general face model and possible instances of the object within the image. The face-detection module is a part of the multimodal biometric-authentication system BioID, described by Frischolz and Werner (2003). Using multimedia data mining in multibiometric systems makes them more reliable due to the presence of an independent piece of a human's trait information.

An example of a practical multimedia data-mining application for medical image data mining is

presented by Mazurkiewicz and Krawczyk (2002). They have used the image data-mining approach to formulate recommendation rules that help physicians to recognize gastroenterological diseases during medical examinations. A parallel environment for image data mining contains a pattern database. Each pattern in the database, considered a representative case, contains formalised text, numeric values, and an endoscopy image. During a patient examination, the automatic classification of the examined case is performed. The system was installed in the Medical Academy of Gdansk, and initial testing results confirm its suitability for further developing.

Multimedia data mining can be applied for discovering structures in video news to extract topics of a sequence or persons involved in the video. A basic approach of multimedia data mining presented by Detyniecki and Marsala (2002) is to separate the visual, audio, and text media channels. The separated multimedia data include features extracted from the video stream, for example, visual spatial content (color, texture, sketch, shape) and visual temporal content (camera or object motion) from the audio stream (loudness, frequency, timbre) and from the text information appearing on the screen. They focused on key-frame color mining in order to notice the appearance of important information on the screen, and on discovering the presence of inlays in a key frame.

Herena et al. (2003) present a multimedia database project called CAESAR™(Civilian American and European Surface Anthropometry Resource Project) and a multimedia data-mining system called Cleopatra, which is intended for utilization by the apparel and transportation industries. The former project consists of anthropometrical and statistical databases that contain data about the worldwide population, 3-D scans of individuals' bodies, consumer habits, lifestyles, and so forth. In the project Cleopatra, a clustering data-mining technique is used to find similar individuals within the population based on an archetype, that is, a typical, real individual within the cluster (see http://www.cleopatra.nrc.ca).

Chen, Shyu, Chen, and Chengcui (2004) propose a framework that uses data mining combined with multimodal processing in extracting the soccer-goal events from soccer videos. It is composed of three major components, namely, video parsing, data prefiltering, and data mining. The integration of data mining and multimodal processing of video is a powerful approach for effective and efficient extraction of soccer-goal events.

MULTIMEDIA DATA-MINING CRITICAL ISSUES

Multimedia data mining can open new threats to informational privacy and information security if not used properly. These activities can give occasion for new types of privacy invasion that may be achieved through the use of cyberspace technology for such things as dataveillance, that is, surveillance by track-

Table 1. A list of multimedia data-mining application domains

• audio analysis (classifying audio track, music mining)
• medical image mining (mammography mining, finding associations between structures and functions of the human brain, formulating recommendation rules for endoscopy recommendation systems)
• mining multimedia data available on the Internet
• mining anthropometry data for the apparel and transportation industries
• movie data mining (movie content analysis, automated rating, getting the story contained in movies)
• pattern recognition (fingerprints, bioinformatics, printed circuit-board inspection)
• satellite-image mining (discovering patterns in global climate change, identifying sky objects, detecting oil spills)
• security (monitoring systems, detecting suspicious customer behaviour, traffic monitoring, outlier detection, multibiometric systems)
• spatiotemporal multimedia-stream data mining (GPS [Global Positioning System], weather forecasting)
• text extracting, segmenting and recognizing from multimedia data
• TV data mining (monitoring TV news, retrieving interesting stories, extracting face sequences from video sequences, extracting soccer-goal events).

ing data shadows that are left behind as individuals undertake their various electronic transactions (Jefferies, 2000). Further invasion can also be occasioned by secondary usage of data that individuals are highly unlikely to be aware of.

Moreover, multimedia data mining is currently still immature. As said by Zhang et al. (2001a, p.18), "The current images association rule mining are far from mature and perfection." Multimedia data are mostly mined separately (Detyniecki & Marsala, 2002). Even if some standards used for multimedia data look very promising, it is too early to draw a conclusion about their usefulness in data mining.

In multimedia data, rare objects are often of great interest. These objects are much harder to identify than common objects. Weiss (2004) states that most data-mining algorithms have a great deal of difficulty dealing with rarity.

CONCLUSION

This article investigates some important issues of multimedia data mining. It presents a short overview of data-mining goals, methods, and techniques; it gives the advantages offered by multimedia data mining and examples of practical applications, application domains, and critical issues; and summarizes the main multimedia data-mining advantages and disadvantages.

Research on text or image mining carried out separately cannot be considered as multimedia data mining unless these media are combined. Multimedia research during the past decade has focused an audio and video media, but now, the wider use of multimodal interfaces and the collection of smart devices with embedded computers should generate a flood of multimedia data, from which knowledge will be extracted using multimedia data-mining methods. The social impact of multimedia data mining is also very important. New threats to informational privacy and information security can occur if these tools are not used properly.

The investigations on multimedia data-mining methods, algorithms, frameworks, and standards should have an impact on the future research in this promising field of information technology. In the future, the author expects that the framework will be made more robust and scalable to a distributed multimedia environment. Other interesting future work concerns multimedia data-mining standardization. Also, the systems need to be evaluated against mining with rarity and the testing of appropriate evaluation metrics. Finally, multimedia data-mining implementations need to be integrated with intelligent user interfaces.

REFERENCES

Bock, H. (2002). The goal of classification. In W. Klosgen & J. M. Zytkow (Eds.), *Handbook of data mining and knowledge discovery* (pp. 254-258). New York: Oxford University Press.

Chen, S. C., Shyu, M. L., Chen, M., & Chengcui, Z. (2004). A decision tree-based multimodal data mining framework for soccer goal detection. *IEEE International Conference on Multimedia and Expo (ICME 2004)*, Taipei, Taiwan.

Ciaccia, P., & Patella, M. (2002). Searching in metric spaces with user-defined and approximate

Table 2. A list of multimedia data-mining advantages and disadvantages

Advantages	Disadvantages
• outlier detection in multimedia	• current algorithms and frameworks are far from being mature and perfect
• extracting and tracking faces and gestures from video	• limited success in specific applications
• understanding and indexing large multimedia files	• lack of multimedia data-mining standards
• possibility to retrieve images by color, texture, and shape	• difficulty dealing with rarity

distances. *ACM Transactions on Database Systems, 27*(4), 398-437.

Detyniecki, M., & Marsala, C. (2002). Fuzzy multimedia mining applied to video news. *The 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2002*, Annecy, France, July 1-5, pp. 1001-1008.

Doorn, M., & de Vries, A. (2000). The psychology of multimedia databases. *Proceedings of the Fifth ACM Conference on Digital Libraries*, 1-9.

Fagin, R., & Wimmers, E. (1997). Incorporating user preferences in multimedia queries. In *Lecture notes in computer science (LNCS): Vol. 1186. Proceedings of the International Conference on Database Theory (ICDT)* (pp. 247-261). Berlin, Heidelberg: Springer-Verlag.

Fayyad, U. (2001). The digital physics of data mining. *Communications of the ACM, 44*(3), 62-65.

Fayyad, U., & Uthurusamy, R. (2002). Evolving data mining into solution for insight. *Communications of the ACM, 45*(8), 28-31.

Frischholz, R. W., & Werner, A. (2003). Avoiding replay-attacks in a face recognition system using head-pose estimation. *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG'03)*, 1-2.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques.* San Mateo, CA: Morgan Kaufmann.

Herena, V., Paquet E., & le Roux, G. (2003). Cooperative learning and virtual reality-based visualization for data mining. In J. Wang (Ed.), *Data mining: Opportunities and challenges* (pp. 55-79). Hershey, PA: Idea Group Publishing.

Hsu, J. (2003). Critical and future trends in data mining: A review of key data mining technologies/applications. In J. Wang (Ed.), *Data mining: Opportunities and challenges* (pp. 437-452). Hershey, PA: Idea Group Publishing.

Jain, A. K., Duin, R. P., & Mao, J. (2000). *Statistical pattern recognition: A review.* Michigan State University Technical Reports, MSU-CSE-00-5. Retrieved on April 4, 2005, from *http://www.cse.msu.edu/cgi-user/web/tech/document?ID=439*

Jain, A. K., & Ross, A. (2004). Multibiometric systems. *Communications of the ACM, 47*(1), 34-40.

Jefferies, P. (2000). Multimedia, cyberspace & ethics. *Proceedings of the IEEE International Conference on Information Visualization (IV'00)*, 99-104.

Jesorsky, O., Kirchberg, K. J., & Frischholz, R. W. (2001). Robust face detection using the Hausdorff distance. In *Lecture notes in computer science (LNCS): Vol. 2091. Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication* (pp. 90-95). Heidelberg: Springer-Verlag.

Kantardzic, M. (2003). *Data mining: Concepts, models, methods, and algorithms.* New York: Wiley-IEEE Press.

Kossmann, D. (2000). The state of the art in distributed query processing. *ACM Computing Surveys, 32*(4), 422-469.

Liautaud, B. (2001). *E-business intelligence turning information into knowledge into profit.* New York: McGraw-Hill.

Mannila, H. (2002). Association rules. In W. Klosgen & J. M. Zytkow (Eds.), *Handbook of data mining and knowledge discovery* (pp. 344-348). New York: Oxford University Press.

Marsala, C. (2000). Fuzzy decision trees to help flexible querying. *Kybernetika, 36*(6), 689-705.

Mazurkiewicz, A., & Krawczyk, H. (2002). A parallel environment for image data mining. *Proceedings of the International Conference on Parallel Computing in Electrical Engineering (PARELEC '02)*, Warsaw, Poland.

Melton, J., & Eisenberg, A. (2001). SQL multimedia and application packages (SQL/MM). *SIGMOD Record, 30*(4), 97-102.

Noirhomme-Fraiture, M. (2000). Multimedia support for complex multidimensional data mining. *Pro-*

ceedings of the First International Workshop on Multimedia Data Mining (MDM/KDD 2000) in conjunction with the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 2000, Boston, MA.

Oliviera, S. R. M., & Zaiane, O. R. (2004). Toward standardization in privacy-preserving data mining. In R. Grossman (Ed.), *Proceedings of the Second International Workshop on Data Mining Standards, Services and Platforms* (pp. 7-17), Seattle, USA. Retrieved on April 4, 2005, from *http://www.cs.ualberta.ca/~zaiane/postscript/dmssp04.pdf*

Pagani, M. (2003). *Multimedia and interactive digital TV: Managing the opportunities created by digital convergence.* Hershey, PA: IRM Press.

Quinlan, J. R. (1986). Introduction of decision trees. *Machine Learning, 1*(1), 86-106.

Rowe, L. A., & Jain, R. (2005). ACM SIGMM retreat report on future directions in multimedia research. *ACM Transactions on Multimedia Computing, Communications and Applications, 1*(1), February, 3-13.

Skowron, A., & Swinarski, R. W. (2004). Information granulation and pattern recognition. In S. K. Pal, L. Polkowski, & A. Skowron (Eds.), *Rough-neural computing* (pp. 599-636). Berlin, Heidelberg: Springer-Verlag.

Swierzowicz, J. (2002). Decision support system for data and Web mining tools selection. In M. Khosrow-Pour (Ed.), *Issues and trends of information technology management in contemporary organizations* (pp. 1118-1120). Hershey, PA: Idea Group Publishing.

Thuraisingham, B. (2002). *XML databases and the semantic Web.* Boca Raton: CRC Press.

Weiss, G. M. (2004). Mining with rarity: A unifying framework. *SIGKDD Explorations, 6*(1), 7-19.

Wijesekera, D., & Barbara, D. (2002). Multimedia applications. In W. Klosgen & J. M. Zytkow (Eds.), *Handbook of data mining and knowledge discovery* (pp. 758-769). New York: Oxford University Press.

Zaiane, O. R., Han, J., Li, Z., & Hou, J. (1998). Mining multimedia data. *Proceedings of Meeting of Minds, CASCON'98*, 1-18.

Zaiane, O. R., Han, J., & Zhu, H. (2000). Mining recurrent items in multimedia with progressive resolution refinement. *Proceedings of the International Conference on Data Engineering ICDE'00*, 15-28.

Zhang, J., Hsu, W., & Lee, L. M. (2001a). Image mining: Issues, frameworks and techniques. In O. R. Zaiane & S. J. Simoff (Eds.), *Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD 2001) in conjunction with Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 2001* (pp. 13-21). San Francisco: ACM Press.

Zhang, J., Hsu, W., & Lee, L. M. (2001b). An information driven framework for image mining. *Proceedings of 12th International Conference on Database and Expert Systems Applications (DEXA),* Munich, Germany.

KEY TERMS

Association Rules: Uncovering interesting trends, patterns, and rules in large data sets with support, *s*, and confidence, *c*.

Confidence: A parameter used in the association-rules method for determining the percent of data cases that support the antecedent of the rule *X* that also support the consequent of the rule *Y* in the set of data cases *D*.

CRISP-DM (Cross-Industry Standard Process for Data Mining): An initiative for standardizing the knowledge-discovery and data-mining process.

Data Mining: An intelligent and automatic process of identifying and discovering useful structures such as patterns, models, and relations in data.

Dataveillance: Surveillance by tracking shadows of data that are left behind as people undertake their electronic transactions.

Image Classification: Classifying a new image, according to the image content, to one of the predefined classes of images (supervised classification).

Image Clustering: Classifying a new image into an image cluster according to the image content (e.g., color, texture, shape, or their combination) without a priori knowledge (unsupervised classification).

Image Indexing: Fast and efficient mechanism based on dimension reduction and similarity measures.

Image Mining: Extracting image patterns, not explicitly stored in images, from a large collection of images.

Image Retrieval: Retrieving an image according to some primitive (e.g., color, texture, shape of image elements) or compound specifications (e.g., objects, given type, abstract attributes).

Isochronous: Processing must occur at regular time intervals.

Key Frame: Representative image of each shot.

Knowledge Discovery in Databases: A process of producing statements that describe objects, concepts, and regularities. It consists of several steps, for example, identification of a problem, cleaning, preprocessing and transforming data, applying suitable data-mining models and algorithms, interpreting, visualizing, testing, and verifying results.

MP3: MPEG audio coding standard layer 3. The main tool for Internet audio delivery.

MPEG: Motion Picture Engineering Group.

MPEG-4: Provides the standardized technological elements enabling the integration of the production, distribution and content access paradigms of digital television, interactive graphics applications and interactive multimedia.

MPEG-7: Multimedia Content Description Interface.

Multimedia Data Mining: Extracting interesting knowledge out of correlated data contained in audio, video, speech, and images.

Object Recognition: A supervised labeling problem based on models of known objects.

Quality of Service: Allocation of resources to provide a specified level of service.

Real Time: Processing must respond within a bounded time to an event.

Shot: A sequence of images in which there is no change of camera.

Support: A parameter used in the association-rules method for determining the percent of data cases that support both the antecedent of the rule X and the consequent of the rule Y in the set of data cases D.

X3D: Open Standards XML (Extensible Markup Language) enabling 3D (dimensional) file format, real-time communication of 3D data across all applications and network applications.

Multimedia Information Design for Mobile Devices

Mohamed Ally
Athabasca University, Canada

INTRODUCTION

There is a rapid increase in the use of mobile devices such as cell phones, tablet PCs, personal digital assistants, Web pads, and palmtop computers by the younger generation and individuals in business, education, industry, and society. As a result, there will be more access of information and learning materials from anywhere and at anytime using these mobile devices. The trend in society today is learning and working on the go and from anywhere rather than having to be at a specific location to learn and work. Also, there is a trend toward ubiquitous computing, where computing devices are invisible to the users because of wireless connectivity of mobile devices. The challenge for designers is how to develop multimedia materials for access and display on mobile devices and how to develop user interaction strategies on these devices. Also, designers of multimedia materials for mobile devices must use strategies to reduce the user mental workload when using the devices in order to leave enough mental capacity to maximize deep processing of the information. According to O'Malley et al. (2003), effective methods for presenting information on these mobile devices and the pedagogy of mobile learning have yet to be developed. Recent projects have started research on how to design and use mobile devices in the schools and in society. For example, the MOBILearn project is looking at pedagogical models and guidelines for mobile devices to improve access of information by individuals (MOBILearn, 2004). This paper will present psychological theories for designing multimedia materials for mobile devices and will discuss guidelines for designing information for mobile devices. The paper then will conclude with emerging trends in the use of mobile devices.

BENEFITS AND LIMITATIONS OF MOBILE DEVICES

There are many benefits of using mobile devices in the workplace, education, and society. In mobile learning (m-learning), users can access information and learning materials from anywhere and at anytime. There are many definitions of m-learning in the field. M-learning is the use of electronic learning materials with built-in learning strategies for delivery on mobile computing devices to allow access from anywhere and at anytime (Ally, 2004a). Another definition of m-learning is any sort of learning that happens when the learner is not at a fixed, predetermined location, or learning that happens when the learner takes advantage of the learning opportunities offered by mobile technologies (O'Malley et al., 2003). With the use of wireless technology, mobile devices do not have to be physically connected to networks in order to access information. Mobile devices are small enough to be portable, which allows users to take the device to any location to access information or learning materials. Because of the wireless connectivity of mobile devices, users can interact with other users from anywhere and at anytime to share information and expertise, complete a task, or work collaboratively on a project. Mobile devices have many benefits, because they allow for mobility while learning and working; however, there are some limitations of mobile devices that designers must be aware of when designing multimedia materials for delivery on mobile devices.

Some of the limitations of mobile devices in delivering multimedia materials include the small screen size for output of information, small input devices, low bandwidth, and challenges when navi-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

gating through the information (Ahonen et al., 2003). Designers of information and learning materials have to be aware of the limited screen size and input device when designing for usability. For example, rather than scrolling for more information on the screen, users of mobile devices must be able to go directly to the information and move back and forth with ease. Information should be targeted to the users' needs when they need it and should be presented efficiently to maximize the display on the mobile device. To compensate for the small screen size of mobile devices, multimedia materials must use rich media to convey the message to the user. For example, rather than present information in textual format, graphics and pictures can be used in such a way to convey the message using the least amount of text. For complex graphics, a general outline of the graphic should be presented on one screen with navigation tools to allow the user to see the details of the graphic on other screens. To present procedures and real-life situations, video clips can be used to present real-life simulations to the user. Also, the interface must be appropriate for individual users and the software system should be able to customize the interface based on individual users' characteristics. When developing multimedia materials for mobile devices, designers must be aware of psychological theories in order to guide the design.

PSYCHOLOGICAL THEORY FOR DEVELOPING MULTIMEDIA MATERIALS FOR MOBILE DEVICES

According to cognitive psychology, learning is an internal process, and the amount learned depends on the processing capacity of the user, the amount of effort expended during the learning process, the quality of the processing, and the user's existing knowledge structure (Ausubel, 1974). These have implications for how multimedia materials should be designed for mobile devices. Designers must include strategies that allow the user to activate existing cognitive structure and conduct quality processing of the information. Mayer et al. (2003) found that when a pedagogical agent was present on the screen as instruction was narrated to students, students who were able to ask questions and receive feedback interactively perform better on a problem-solving transfer test when compared to students who only received on-screen text with no narration. It appears that narration by a pedagogical agent encouraged deep processing, which resulted in higher-level learning. According to Paivio's theory of dual coding, memory is enhanced when information is represented both in verbal and visual forms (Paivio, 1986). Presenting materials in both textual and visual forms will involve more processing, resulting in better storage and integration in memory (Mayer et al., 2004). Tabbers et al. (2004) found that in a Web-based multimedia lesson, students who received visual cues to pictures scored higher on a retention test when compared to students who did not receive the cues for the pictures. Also, strategies can be included to get the user to retrieve existing knowledge to process the information presented. For example, a comparative advance organizer can be used to activate existing knowledge structure to process the incoming information, or an expository advance organizer can be presented and stored in memory to help incorporate the details in the information (Ally, 2004a; Ausubel, 1974).

Constructivism is a theory of learning that postulates that learners are active during the learning process, and that they use their existing knowledge to process and personalize the incoming information. Constructivists claim that learners interpret information and the world according to their personal realities, and that they learn by observation, processing, and interpretation and then personalize the information into their existing knowledge bases (Cooper, 1993). Users learn best when they can contextualize what they learn for immediate application and to acquire personal meaning. According to Sharples (2000), mobile learning devices allow learners to learn wherever they are located and in their personal context so that the learning is meaningful. Also, mobile devices facilitate personalized learning, since learning is contextualized where learning and collaboration can occur from anywhere and anytime. According to constructivism, learners are not passive during the learning process. As a result, interaction on mobile devices must include strategies to actively process and internalize the information. For example, on a remote job site, a user can access the information using a mobile device for just-in-time training and then apply the information right away.

As a result, designers must use instructional strategies to allow users to apply what they learn.

DESIGN GUIDELINES FOR MULTIMEDIA MATERIALS FOR MOBILE DEVICES

Cater for the User of Mobile Devices

- **Design for the User:** One of the variables that designers tend to ignore when they develop multimedia materials for mobile devices is the user of the devices. Different users have different learning styles; some users may be visual, while others may be verbal (Mayer & Massa, 2003). Users have different learning styles and preferences; strategies must be included and information presented in different ways in order to cater to the different learning styles and preferences (Ally & Fahy, 2002). Graphic overviews can be used to cater to users who prefer to get the big picture before they go into the details of the information. For active learners, information can be presented on the mobile device, and then the user can be given the opportunity to apply the information. For the creative users, there must be opportunities to apply the information in real-life applications so that they go beyond what was presented. The multimedia materials and information have to be designed with the user in mind in order to facilitate access, learning, and comprehension. Also, the user should have control of what he or she wants to access in order to go through the multimedia materials based on preferred learning styles and preferences. For users in remote locations with low bandwidth or limited wireless access, information that takes a long time to download should be redesigned to facilitate efficient download.
- **Adapt the Interface to the User:** An interface is required to coordinate interaction between the user and the information. To compensate for the small screen size of the display of the mobile device, the interface of the mobile device must be designed properly. The interface can be graphical and should present limited information on the screen to prevent information overload in short-term memory. The system should contain intelligent software agents to determine what the user did in the past and to adapt the interface for future interaction with the information. The software system must be proactive by anticipating what the user will do next and must provide the most appropriate interface for the interaction to enhance learning. Users must be able to jump to related information without too much effort. The interface must allow the user to access the information with minimal effort and move back to previous information with ease. For sessions that are information-intense, the system must adjust the interface to prevent information overload. Some ways to prevent information overload include presenting less concepts on one screen or organizing the information in the form of concept maps to give the overall structure of the information and then presenting the details by linking to other screens with the details. The interface also must use good navigational strategies to allow users to move back and forth between displays. Navigation can also be automatic based on the intelligence gathered on the user's current progress and needs.
- **Design for Minimum Input:** Because of the small size of the input device, multimedia materials must be designed to require minimum input from users. Input can use pointing or voice input devices to minimize typing and writing. Because mobile devices allow access of information from anywhere at anytime, the device must have input and output options to prevent distractions when using the mobile devices. For example, if someone is using a mobile device in a remote location, it may be difficult to type on a keyboard or use a pointing device. The mobile technology must allow the user to input data using voice input or touch screen.
- **Build Intelligent Software Agents to Interact with the User:** Intelligent software systems can be built to develop an initial profile of the user and then present materials that will benefit the specific user, based on the user profile. As the intelligent agent interacts with the user, it learns about the user and adapts the format of the information, the interface, and the navigation pattern according to the user's style and

needs. Knowing the user's needs and style will allow the intelligent software system to access additional materials from the Internet and other networks to meet the needs of user (Cook et al., 2004).

- **Use a Personalized Conversational Style:** Multimedia information and learning materials can be presented to the user in a personalized style or a formal style. In a learning situation, information should be presented in a personalized style, since the user of the mobile device may be in a remote location and will find this style more connected and personal. Mayer et al. (2004) found that students who received a personalized version of a narrated animation performed significantly better on a transfer test when compared to students who received a non-personalized, formal version of the narrated animation. They claimed that the results from the study are consistent with the cognitive theory of multimedia learning, where personalization results in students processing the information in an active way, resulting in higher-level learning and transfer to other situations.

Design to Improve the Quality of Information Processing on Mobile Devices

- **Chunk Information for Efficient Processing:** Designers of materials for mobile devices must use presentation strategies to enable users to process the materials efficiently because of the limited display capacity of mobile devices and the limited processing capacity of human working memory. Information should be organized or chunked in segments of appropriate and meaningful size to facilitate processing in working memory. An information session on a mobile device can be seen as consisting of a number of information objects sequenced in a predetermined way or sequenced based on the user needs. Information and learning materials for mobile devices should take the form of information and learning objects that are in an electronic format, reusable, and stored in a repository for access anytime and from anywhere (McGreal, 2004). Information objects and learning objects allow for instant assembly of learning materials by users and intelligent software agents to facilitate just-in-time learning and information access. The information can be designed in the form of information objects for different learning styles and characteristics of users (Ally, 2004b). The objects then are tested and placed in an electronic repository for just-in-time access from anywhere and at anytime using mobile devices.
- **Use High-Level Concept Maps to Show Relationships:** A concept map or a network diagram can be used to show the important concepts in the information and the relationship between the concepts rather than present information in a textual format. High-level concept maps and networks can be used to represent information spatially so that students can see the main ideas and their relationships (Novak, Gowin, & Johanse, 1983). Tusack (2004) suggests the use of site maps as the starting point of interaction that users can link back in order to continue with the information or learning session. Eveland et al. (2004) compared linear Web site designs and non-linear Web site designs and reported that linear Web site designs encourage factual learning, while non-linear Web site designs increase knowledge structure density. One can conclude that the non-linear Web site designs show the interconnection of the information on the Web site, resulting in higher-level learning.

EMERGING TRENDS IN DESIGNING MULTIMEDIA MATERIALS FOR MOBILE DEVICES

The use of mobile devices with wireless technology allow access of information and multimedia materials from anywhere and anytime and will dramatically alter the way we work and conduct business and how we interact with each other (Gorlenko & Merrick, 2003). For example, mobile devices can make use of Global Positioning Systems to determine where users are located and connect them with users in the same location so that they can work collaboratively on projects and learning materials. There will be exponential growth in the use of mobile devices to access information and learning materials, since the cost of the devices will be lower than desktop com-

puters, and users can access information from anywhere and at anytime. Also, the use of wireless mobile devices would be more economical, since it does not require the building of the infrastructure to wire buildings. The challenge for designers of multimedia materials for mobile devices is how to standardize the design for use by different types of devices. Intelligent software agents should be built into mobile devices so that most of the work is done behind the scenes, minimizing input from users and the amount of information presented on the display of the mobile devices. Because mobile devices provide the capability to access information from anywhere, future multimedia materials must be designed for international users.

CONCLUSION

In the past, the development of multimedia materials and mobile devices concentrated on the technology rather than the user. Future development of multimedia materials for mobile devices should concentrate on the user to drive the development and delivery (Gorlenko & Merrick, 2003). Mobile devices can be used to deliver information and learning materials to users, but the materials must be designed properly in order to compensate for the small screen of the devices and the limited processing and storage capacity of a user's working memory. Learning materials need to use multimedia strategies that are information-rich rather than mostly text.

REFERENCES

Ahonen, M., Joyce, B., Leino, M., & Turunen, H. (2003). Mobile learning: A different viewpoint. In H. Kynaslahti, & P. Seppala (Eds.), *Mobile learning* (pp. 29-39). Finland: Edita Publishing Inc.

Ally, M. (2004a). Using learning theories to design instruction for mobile learning devices. *Proceedings of the Mobile Learning 2004 International Conference*, Rome.

Ally, M. (2004b). Designing effective learning objects for distance education. In R. McGreal (Ed.), *Online education using learning objects* (pp. 87-97). London: RoutledgeFalmer.

Ally, M., & Fahy, P. (2002). Using students' learning styles to provide support in distance education. *Proceedings of the Eighteenth Annual Conference on Distance Teaching and Learning*, Madison, Wisconsin.

Ausubel, D.P. (1974). *Educational psychology: A cognitive view*. New York: Holt, Rinehart and Winston.

Cook, D.J., Huber, M., Yerraballi, R., & Holder, L.B. (2004). Enhancing computer science education with a wireless intelligent simulation environment. *Journal of Computing in Higher Education*, *16*(1), 106-127.

Cooper, P.A. (1993). Paradigm shifts in designing instruction: From behaviorism to cognitivism to constructivism. *Educational Technology*, *33*(5), 12-19.

Eveland, W.P., Cortese, J., Park, H., & Dunwoody, S. (2004). How website organization influences free recall, factual knowledge, and knowledge structure density. *Human Communication Research*, *30*(2), 208-233.

Gorlenko, L., & Merrick, R. (2003). No wires attached: Usability challenges in the connected mobile world. *IBM Systems Journal*, *42*(4), 639-651.

Mayer, R.E., Dow, T.D., & Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds. *Journal of Educational Psychology*, *95*(4), 806-813.

Mayer, R.E., Fennell, S., Farmer, L., & Campbell, J. (2004). A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style. *Journal of Educational Psychology*, *96*(2), 389-395.

Mayer, R.E., & Massa, L.J. (2003). Three facets of visual and verbal learners: Cognitive ability, cognitive style, and learning preference. *Journal of Educational Psychology*, *95*(4), 833-846.

McGreal, R. (2004). *Online education using learning objects*. London: Routledge/Falmer.

MOBILearn Leaflet. (2004): Next-generation paradigms and interfaces for technology supported learning in a mobile environment exploring the potential of

ambient intelligence. Retrieved September 8, 2004, from *http://www.mobilearn.org/results/results.htm*

Novak, J.D., Gowin, D.B., & Johanse, G.T. (1983). The use of concept mapping and knowledge vee mapping with junior high school science students. *Science Education, 67,* 625-645.

O'Malley, C., et al. (2003). Guidelines for learning/teaching/tutoring in a mobile environment. Retrieved September 8, 2004, from *http://www.mobilearn.org/results/results.htm*

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford: Oxford University Press.

Sharples, M. (2000). The design of personal mobile technologies for lifelong learning. *Computers and Education*, *34*, 177-193.

Tabbers, H.K., Martens, R.L., & van Merrienboer, J.J.G. (2004). Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British Journal of Educational Psychology*, *74*, 71-81.

Tusack, K. (2004). Designing Web pages for handheld devices. *Proceedings of the 20th Annual Conference on Distance Teaching and Learning*, Madison, Wisconsin.

KEY TERMS

Advance Organizer: A general statement at the beginning of the information or lesson to activate existing cognitive structure or to provide the appropriate cognitive structure to learn the details in the information or the lesson.

Concept Map: A graphic outline that shows the main concepts in the information and the relationship between the concepts.

Intelligent Software Agent: A computer application software that is proactive and capable of flexible autonomous action in order to meet its design objectives set out by the designer. The software learns about the user and adapts the interface and the information to the user's needs and style.

Interface: The components of the computer program that allow the user to interact with the information.

Learning Object: A digital resource that is stored in a repository that can be used and reused to achieve a specific learning outcome or multiple outcomes (Ally, 2004b).

Learning Style: A person's preferred way to learn and process information, interact with others, and complete practical tasks.

Mobile Device: A device that can be used to access information and learning materials from anywhere and at anytime. The device consists of an input mechanism, processing capability, storage medium, and display mechanism.

Mobile Learning (M-Learning): Electronic learning materials with built-in learning strategies for delivery on mobile computing devices to allow access from anywhere and at anytime.

Multimedia: A combination of two or more media to present information to users.

Short-Term Memory: The place where information is processed before the information is transferred to long-term memory. The duration of short-term memory is very short, so information must be processed efficiently to maximize transfer to long-term memory.

Ubiquitous Computing: Computing technology that is invisible to the user because of wireless connectivity and transparent user interface.

User: An individual who interacts with a computer system to complete a task, learn specific knowledge or skills, or access information.

Wearable Computing Devices: Devices that are attached to the human body so that the hands are free to complete other tasks.

Multimedia Information Retrieval at a Crossroad

Qing Li
City University of Hong Kong, China

Jun Yang
Carnegie Mellon University, USA

Yueting Zhuang
Zhejiang University, China

INTRODUCTION

In the late 1990s, the availability of powerful computing capability, large storage devices, high-speed networking and especially the advent of the Internet, led to a phenomenal growth of digital multimedia content in terms of size, diversity and impact. As suggested by its name, "multimedia" is a name given to a collection of multiple types of data, which include not only "traditional multimedia" such as images and videos, but also emerging media such as 3D graphics (like VRML objects) and Web animations (like Flash animations). Furthermore, multimedia techniques have been penetrating into a growing number of applications, ranging from document-editing software to digital libraries and many Web applications. For example, most people who have used Microsoft Word have tried to insert pictures and diagrams into their documents, and they have the experience of watching online video clips, such as movie trailers. In other words, multimedia data have been in every corner of the digital world. With the huge volume of multimedia data, finding and accessing the multimedia documents that satisfy people's needs in an accurate and efficient manner became a non-trivial problem. This problem is defined as multimedia information retrieval.

The core of multimedia information retrieval is to compute the degree of relevance between users' information needs and multimedia data. A user's information need is expressed as a query, which can be in various forms, such as a line of free text like, "Find me the photos of George Washington"; a few key words, like, "George Washington photo"; or a media object, like a picture of George Washington. Moreover, the multimedia data are also represented by a certain form of summarization, typically called an index, which is directly matched against queries. Similar to a query, the index can take a variety of forms, including key words and features such as color histograms and motion vectors, depending on the data and task characteristics.

For textual documents, mature information retrieval (IR) technologies have been developed and successfully applied in commercial systems such as Web search engines. In comparison, the research on multimedia retrieval is still in its early stage. Unlike textual data, which can be well represented by key words as an index, multimedia data lack an effective, semantic-level representation (or index) that can be computed automatically, which makes multimedia retrieval a much harder research problem. On the other hand, the diversity and complexity of multimedia offer new opportunities for its retrieval task to be leveraged by the state of the art in various research areas. In fact, research on multimedia retrieval has been initiated and investigated by researchers from areas of multimedia database, computer vision, natural language processing, human-computer interaction and so forth. Overall, it is currently a very active research area that has many interactions with other areas.

In the following sections, we will overview the techniques for multimedia information retrieval and review the applications and challenges in this area. Then, future trends will be discussed. Some important terms in this area are defined at the end of this article.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

MULTIMEDIA RETRIEVAL TECHNIQUES

Despite the various techniques proposed in literature, there exist two major approaches to multimedia retrieval; namely, text-based and content-based. Their main difference lies in the type of index: The former approach uses text (key words) as the index, whereas the latter uses low-level features extracted from multimedia data. As a result, they differ from each other in many other aspects, ranging from feature extraction to similarity measurement.

Text-Based Multimedia Retrieval

Text-based multimedia retrieval approaches apply mature IR techniques to the domain of multimedia retrieval. A typical text-IR method matches text queries posed by users with descriptive key words extracted from documents. To use the method for multimedia, textual descriptions (typically key word annotations) of the multimedia objects need to be extracted. Once the textual descriptions are available, multimedia retrieval boils down to a text-IR problem. In early years, such descriptions were usually obtained by manually annotating the multimedia data with key words (Tamura & Yokoya, 1984). Apparently, this approach is not scalable to large datasets, due to its labor-intensive nature and vulnerability to human biases. There also have been proposals from computer vision and pattern recognition areas on automatically annotating the images and videos with key words based on their low-level visual/audio features (Barnard, Duygulu, Freitas, Forsyth, Blei, D. & Jordan, 2003). Most of these approaches involve supervised or unsupervised machine learning, which tries to map low-level features into descriptive key words. However, due to the large gap between the multimedia data form (e.g., pixels, digits) and their semantic meanings, it is unlikely to produce high-quality key word annotations automatically. Some of the systems are semi-automatic, attempting to propagate key words from a set of initially annotated objects to other objects. In other applications, descriptive key words can be easily accessible for multimedia data. For example, for images and videos embedded in Web pages, the text surrounding them is usually a good description, which has been explored in the work of Smith and Chang (1997).

Since key word annotations can precisely describe the semantic meanings of multimedia data, the text-based retrieval approach is effective in terms of retrieving multimedia data that are *semantically relevant* to the users' needs. Moreover, because many people find it convenient and effective to use text (or key words) to express their information requests, as demonstrated by the fact that most commercial search engines (e.g., Google) support text queries, this approach has the advantage of being amenable to average users. But the bottleneck of this approach is still on the acquisition of key word annotations, since there are no indexing techniques that guarantee both efficiency and accuracy if the annotations are not directly available.

Content-Based Multimedia Retrieval

The idea of content-based retrieval first came from the area of content-based image retrieval (CBIR) (Flickner, Sawhney, Niblack, Ashley, Huang, Dom, Gorkani, Hafner, Lee, Petkovic, Steele & Yanker, 1995; Smeulders, Worring, Santini, Gupta & Jain, 2000). Gradually, the idea has been applied to retrieval tasks for other media types, resulting in content-based video retrieval (Hauptmann et al., 2002; Somliar, 1994) and content-based audio retrieval (Foote, 1999). The word "content" here refers to the bottom-level representation of the data, such as pixels for bitmap images, MPEG bit-streams for MPEG-format video and so forth. Content-based retrieval, as opposed to a text-based one, exploits the features that are (automatically) extracted from the low-level representation of the data, usually denoted as low-level features since they do not directly capture the high-level meanings of the data. (In a sense, text-based retrieval of documents is also "content based," since key words are extracted from the content of documents.) The low-level features used for retrieval depend on the specific data type: A color histogram is a typical feature for image retrieval, motion vector is used for video retrieval, and so forth. Despite the heterogeneity of the features, in most cases, they can be transformed into feature vector(s). Thus, the similarity between media objects can be measured by the distance of their respective feature vectors in the vector space under certain distance metrics. Various distance measures, such as Euclidean distance and

M-distance, can be used as the similarity metrics. This has a correspondence to the vector-based model for (text) information retrieval, where a bag of key words is also represented as a vector.

Content-based retrieval also influences the way a query is composed. Since a media object is represented by its low-level feature vector(s), a query must be also transformed into a feature vector to match against the object. This results in query-by-example (QBE) (Flickner et al., 1995), a new search paradigm where media objects such as images or video clips are used as query examples to find other objects similar to them, where "similar" is defined mainly at perceptual levels (i.e., looks like or sounds like). In this case, feature vector(s) extracted from the example object(s) are matched with the feature vectors of the candidate objects. A vast majority of content-based retrieval systems use QBE as its search paradigm. However, there are also content-based systems that use alternative ways to let users specify their intended low-level features, such as by selecting from some templates or a small set of feature options (i.e., "red," "black" or "blue").

The features and similarity metrics used by many content-based retrieval systems are chosen heuristically and are therefore ad-hoc and unjustified. It is very questionable that the features and metrics are optimal or close to optimal. Thus, there have been efforts seeking for theoretically justified retrieval approaches whose optimality is guaranteed under certain circumstances. Many of these approaches treat retrieval as a machine-learning problem of finding the most effective (weighted) combination of features and similarity metrics to solve a particular query or set of queries. Such learning can be done online in the middle of the retrieval process, based on user-given feedback evaluations or automatically derived "pseudo" feedback. In fact, relevance feedback (Rui, Huang, Ortega & Mehrotra, 1998) has been one of the hot topics in content-based retrieval. Off-line learning has also been used to find effective features/weights based on previous retrieval experiences. However, machine learning is unlikely to be the magic answer for the content-based retrieval problem, because it is impossible to have training data for basically an infinite number of queries, and users are usually unwilling to give feedback.

Overall, content-based retrieval has the advantage of being fully automatic from the feature extraction to similarity computation, and thus scalable to real systems. With the QBE search paradigm, it is also able to capture the perceptual aspects of multimedia data that cannot be easily depicted by text. The downside of content-based retrieval is mainly due to the so-called "semantic gap" between low-level features and the semantic meanings of the data. Given that users prefer semantically relevant results, content-based methods suffer from the low precision/recall problem, which prevents them from being used in commercial systems. Another problem lies in the difficulty of finding a suitable example object to form an effective query if the QBE paradigm is used.

APPLICATIONS AND CHALLENGES

Though far from mature, multimedia retrieval techniques have been widely used in a number of applications. The most visible application is on Web search engines for images, such as the Google Image search engine (Brin & Page, 1998), Ditto.com and so forth. All these systems are text-based, implying that a text query is a better vehicle of users' information need than an example-based query. Content-based retrieval is not applicable here due to its low accuracy problem, which gets even worse due to the huge data volume. Web search engines acquire textual annotations (of images) automatically by analyzing the text in Web pages, but the results for some popular queries may be manually crafted. Because of the huge data volume on the Web, the relevant data to a given query can be enormous. Therefore, the search engines need to deal with the problem of "authoritativeness" – namely, determining how authoritative a piece of data is – besides the problem of relevance. In addition to the Web, there are many digital libraries, such as Microsoft Encarta Encyclopedia, that have the facilities for searching multimedia objects like images and video clips by text. The search is usually realized by matching manual annotations with text queries.

Multimedia retrieval techniques have also been applied to some narrow domains, such as news videos, sports videos and medical imaging. NIST TREC Video Retrieval Evaluation has attracted many research efforts devoted to various retrieval tasks on broadcast news video based on automatic analysis of video content. Sports videos, like basketball

programs and baseball programs, have been studied to support intelligent access and summarization (Zhang & Chang, 2002). In the medical imaging area, for example, Liu et al. (2002) applied retrieval techniques to detect a brain tumor from CT/MR images. Content-based techniques have achieved some level of success in these domains, because the data size is relatively small, and domain-specific features can be crafted to capture the idiosyncrasy of the data. Generally speaking, however, there is no killer application where content-based retrieval techniques can achieve a fundamental breakthrough.

The emerging applications of multimedia also raise new challenges for multimedia retrieval technologies. One such challenge comes from the new media formats emerged in recent years, such as Flash animation, PowerPoint file and Synchronized Multimedia Integration Language (SMIL). These new formats demand specific retrieval methods. Moreover, their intrinsic complexity (some of them can recursively contain media components) brings up new research problems not addressed by current techniques. There already have been recent efforts devoted to these new media, such as Flash animation retrieval (Yang, Li, Liu & Zhuang, 2002a) and PowerPoint presentation retrieval. Another challenge rises from the idea of retrieving multiple types of media data in a uniform framework, which will be discussed next.

FUTURE TRENDS

In a sense, most existing multimedia retrieval methods are not genuinely for "multimedia," but are for a specific type (or modality) of non-textual data. There is, however, the need to design a real "multimedia" retrieval system that can handle multiple data modalities in a cooperative framework. First, in multimedia databases like the Web, different types of media objects coexist as an organic whole to convey the intended information. Naturally, users would be interested in seeing the complete information by accessing all the relevant media objects regardless of their modality, preferably from a single query. For example, a user interested in a new car model would like to see pictures of the car and meanwhile read articles on it. Sometimes, depending on the physical conditions, such as networks and displaying devices, users may want to see a particular presentation of the information in appropriate modality(-ies). Furthermore, some data types, such as video, intrinsically consist of data of multiple modalities (audio, closed-caption, video images). It is advantageous to explore all these modalities and let them complement each other to obtain a better retrieval effect. To sum, a retrieval system that goes across different media types and integrates multi-modality information is highly desirable.

Informedia (Hauptmann et al., 2002) is a well-known video retrieval system that successfully combines multi-modal features. Its retrieval function not only relies on the transcript generated from a speech recognizer and/or detected from overlaid text on screen, but also utilizes features such as face detection and recognition results, image similarity and so forth. Statistical learning methods are widely used in Informedia to intelligently combine the various types of information. Many other systems integrate features from at least two modalities for retrieval purpose. For example, the WebSEEK system (Smith & Chang, 1997) extracts key words from the surrounding text of images and videos in Web pages, which is used as their indexes in the retrieval process. Although the systems involve more than one media type, typically, textual information plays the vital role in providing the (semantic) annotation of the other media types.

Systems featuring a higher degree of integration of multiple modalities are emerging. More recently, MediaNet (Benitez, Smith & Chang, 2002) and multimedia thesaurus (MMT) (Tansley, 1998) are proposed, both of which seek to provide a multimedia representation of a semantic concept – a concept described by various media objects including text, image, video and so forth – and establish the relationships among these concepts. MediaNet extends the notion of relationships to include even perceptual relationships among media objects.

Yang, Li and Zhuang (2002b) propose a very comprehensive and flexible model named *Octopus* to perform an "aggressive" search of multi-modality data. It is based on a multi-faceted knowledge base represented by a layered graph model, which captures the relevance between media objects of any type from various perspectives, such as the similarity on low-level features, structural relationships such as hyperlinks and semantic relevance. Link analysis

techniques can be used to find the most relevant objects for any given object in the graph. This new model can accommodate knowledge from various sources, and it allows a query to be composed flexibly using either text or example objects, or both.

CONCLUSION

Multimedia information retrieval is a relatively new area that has been receiving more attention from various research areas like database, computer vision, natural language and machine learning, as well as from industry. Given the continuing growth of multimedia data, research in this area will expectedly become more active, since it is critical to the success of various multimedia applications. However, technological breakthroughs and killer applications in this area are yet to come, and before that, multimedia retrieval techniques can hardly be migrated to commercial applications. The breakthrough in this area depends on the joint efforts from its related areas, and therefore, it offers researchers opportunities to tackle the problem from different paths and with different methodologies.

REFERENCES

Barnard, K., Duygulu, P., Freitas, N., Forsyth, D., Blei, D., & Jordan, M. (2003). Matching words and pictures. *Journal of Machine Learning Research, 3,* 1107-1135.

Benitez, A.B., Smith, J.R., & Chang, S.F. (2000). MediaNet: A Multimedia Information Network for knowledge representation. *Proceedings of the SPIE 2000 Conference on Internet Multimedia Management Systems*, 4210.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the 7th International World Wide Web Conference,* 107-117.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., & Yanker, P. (1995). Query by image and video content: The QBIC system. *IEEE Computer, 28*(9), 23-32.

Foote, J. (1999). An overview of audio information retrieval. Multimedia Systems, 7(1), 2-10.

Hauptmann, A., et al. (2002). Video classification and retrieval with the Informedia Digital Video Library System. *Text Retrieval Conference (TREC02)*, Gaithersburg, MD.

Liu, Y., Lazar, N., & Rothfus, W. (2002). Semantic-based biomedical image indexing and retrieval. *International Conference on Diagnostic Imaging and Analysis (ICDIA 2002)*.

Lu, Y, Hu, C., Zhu, X., Zhang, H., Yang Q., (2000). A unified framework for semantics and feature based relevance feedback in image retrieval systems. *Proceedings of ACM Multimedia Conference,* 31-38.

NIST TREC Video Retrieval Evaluation. Retrieved from *www-nlpir.nist.gov/projects/trecvid/*

Rui, Y., Huang, T.S., Ortega, M., & Mehrotra, S. (1998). Relevance feedback: A power tool for interactive content-based image retrieval. IEEE Trans on Circuits and Systems for Video Technology *(Special Issue on Segmentation, Description, and Retrieval of Video Content) 8,* 644-655.

Smeulders, M., Worring, S., Santini, A., Gupta, & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1349-1380.

Smith, J.R., & Chang, S.F. (1997). Visually searching the Web for content. *IEEE Multimedia Magazine, 4*(3), 12-20.

Somliar, S.W., Zhang, H., et al. (1994). Content-based video indexing and retrieval. *IEEE MultiMedia, 1*(2), 62-72.

Synchronized Multimedia Integration Language (SMIL). Retrieved from *www.w3.org/AudioVideo/*

Tamura, H., & Yokoya, N. (1984) Image database systems: A survey. *Pattern Recognition, 17*(1), 29-43.

Tansley, R. (1998). The Multimedia Thesaurus: An aid for multimedia information retrieval and navigation (masters thesis). *Computer Science,* University of Southampton.

Yang, J., Li, Q., Liu W., & Zhuang, Y. (2002a). FLAME: A generic framework for content-based Flash retrieval. *ACM MM'2002 Workshop on Multimedia Information Retrieval*, Juan-les-Pins, France.

Yang, J., Li, Q., & Zhuang, Y. (2002b). Octopus: Aggressive search of multi-modality data using multifaceted knowledge base. *Proceedings of the 11th International Conference on World Wide Web*, 54-64.

Zhang, D., & Chang, S.F (2002). Event detection in baseball video using superimposed caption recognition. *Proceedings of ACM Multimedia Conference*, 315-318.

KEY TERMS

Content-Based Retrieval: An important retrieval method for multimedia data, which uses the low-level features (automatically) extracted from the data as the indexes to match with queries. Content-based image retrieval is a good example. The specific low-level features used depend on the data type: Color, shape and texture features are common features for images, while kinetic energy and motion vectors are used to describe video data. Correspondingly, a query also can be represented in terms of features so that it can be matched against the data.

Index: In the area of information retrieval, an "index" is the representation or summarization of a data item used for matching with queries to obtain the similarity between the data and the query, or matching with the indexes of other data items. For example, key words are frequently used indexes of textual documents, and color histogram is a common index of images. Indexes can be manually assigned or automatically extracted. The text description of an image is usually manually given, but its color histogram can be computed by programs.

Information Retrieval (IR): The research area that deals with the storage, indexing, organization of, search, and access to information items, typically textual documents. Although its definition includes multimedia retrieval (since information items can be multimedia), the conventional IR refers to the work on textual documents, including retrieval, classification, clustering, filtering, visualization, summarization and so forth. The research on IR started nearly half a century ago and it grew fast in the past 20 years with the efforts of librarians, information experts, researchers on artificial intelligence and other areas. A system for the retrieval of textual data is an IR system, such as all the commercial Web search engines.

Multimedia Database: A database system dedicated to the storage, management and access of one or more media types, such as text, image, video, sound, diagram and so forth. For example, an image database such as Corel Image Gallery that stores a large number of pictures and allows users to browse them or search them by key words can be regarded as a multimedia database. An electronic encyclopedia such as Microsoft Encarta Encyclopedia, which consists of tens of thousands of multimedia documents with text descriptions, photos, video clips and animations, is another typical example of a multimedia database.

Multimedia Document: A multimedia document is a natural extension of a conventional textual document in the multimedia area. It is defined as a digital document composed of one or multiple media elements of different types (text, image, video, etc.) as a logically coherent unit. A multimedia document can be a single picture or a single MPEG video file, but more often it is a complicated document, such as a Web page, consisting of both text and images.

Multimedia Information Retrieval (System): Storage, indexing, search and delivery of multimedia data such as images, videos, sounds, 3D graphics or their combination. By definition, it includes works on, for example, extracting descriptive features from images, reducing high-dimensional indexes into low-dimensional ones, defining new similarity metrics, efficient delivery of the retrieved data and so forth. Systems that provide all or part of the above functionalities are multimedia retrieval systems. The Google image search engine is a typical example of such a system. A video-on-demand site that allows people to search movies by their titles is another example.

Multi-Modality: Multiple types of media data, or multiple aspects of a data item. Its emphasis is on the existence of more than one type (aspects) of data. For example, a clip of digital broadcast news video has

multiple modalities, include the audio, video frames, closed-caption (text) and so forth.

Query-by-Example (QBE): A method of forming queries that contains one or more media object(s) as examples with the intention of finding similar objects. A typical example of QBE is the function of "See Similar Pages" provided in the Google search engine, which supports finding Web pages similar to a given page. Using an image to search for visually similar images is another good example.

Multimedia Instructional Materials in MIS Classrooms[1]

Randy V. Bradley
Troy University, USA

Victor Mbarika
Southern University and A&M College, USA

Chetan S. Sankar
Auburn University, USA

P.K. Raju
Auburn University, USA

INTRODUCTION

Researchers and major computing associations such as the Association of Information Systems (AIS) and the Association of Computing Machinery (ACM) have invested much effort in the last two decades to shape the information system (IS) curriculum in a way that addresses developments and rapid changes in the IS industry (Gorgone, Gray, Feinstein, Kasper, Luftman, Stohr et al., 2000; Nunamaker, Couger & Davis, 1982). A major objective has been to help overcome the skill shortages that exist in the IS field, a trend that is expected to continue in the years ahead (Gorgone et al., 2000). While there exist a plethora of students joining IS programs around the world (usually for the remunerative promises that goes with an IS degree), students do not seem to gain the kind of knowledge and technical expertise needed to face real-world challenges when they take on positions in the business world. There is, therefore, the need to prepare IS students for real-world challenges by developing their technical and decision-making skills.

The purpose of this article, therefore, is to help IS researchers and educators evaluate the potential of LITEE[2] multimedia instructional materials as a pedagogy that assists instructors in conveying IT concepts to students. Another purpose is to present an instruction manual that includes step-by-step instructions about how to use LITEE multimedia instructional materials in a typical IS introductory class. In addition, we outline the most critical issues that should be considered prior to using multimedia instructional materials developed by LITEE and similar organizations. This article should be especially useful to instructors and administrators who desire to use such multimedia instructional materials in IS undergraduate classrooms.

This article is organized in six sections. Following this introduction, we define multimedia, followed by a discussion of the benefits and limitations of using multimedia instructional materials in IS undergraduate classrooms. Then we offer practical guidance for those using multimedia instructional materials in IS undergraduate classrooms. Next, we suggest evaluating students' performance when using multimedia instructional materials. And finally, we conclude the instruction manual.

DEFINING MULTIMEDIA

The term multimedia generally refers to the combination of several media of communication, such as text, graphics, video, animation, music and sound effects (Gaytan & Slate, 2002, 2003). When used in conjunction with computer technology, multimedia has been referred to by some as interactive media (Fetterman, 1997; Gaytan & Slate, 2002, 2003). Gaytan and Slate cite four components essential to multimedia: (a) a computer to coordinate sound, video and interactivity; (b) hyperlinks that connect the information; (c) navigational tools that browse the Web site or Web page containing the connected

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

information; and (d) methods to gather, process and communicate information and ideas. Multimedia does not exist if one of these four components is missing, and depending upon which component is missing, the product might be referred to by a different name. For example, the product might be referred to as (a) "mixed media" if the component that provides interactivity is missing; (b) a "bookshelf" if it lacks links to connect the information; (c) a "movie" if it lacks navigational tools allowing the user to choose a course of action; and (d) "television" if it does not provide users the opportunity to create and contribute their own ideas (Gaytan & Slate, 2002, 2003). Thus, multimedia, appropriately defined, is "the use of a computer to present and combine text, graphics, audio and video with links and tools that allows the user to navigate, interact, create and communicate" (Gaytan & Slate, 2002, 2003).

BENEFITS AND LIMITATIONS OF USING MULTIMEDIA INSTRUCTIONAL MATERIALS IN UNDERGRADUATE IS CLASSROOMS

Nielsen (1995) reports that multimedia systems enable non-linear access to vast amounts of information. Other researchers show that with multimedia users can explore information in-depth on demand, and interact with instructional materials on a self-paced mode (Barrett, 1988; Collier, 1987). Others state that multimedia is attention-capturing or engaging to use and represents a natural form of representation with respect to the workings of the human mind (Delany & Gilbert, 1991; Jonassen, 1989). Oliver and Omari's (1999) study suggested that while print (paper-based) instructional materials provided a sound means to guide and direct students' use of the World Wide Web (WWW) learning materials, the actual WWW materials were more suited to supporting interactive learning activities rather than conveying content and information. Sankar and Raju (2002) report that multimedia instructional materials produced at their laboratory and used in business classrooms are aimed at both improving what students learn and the way students learn. Thus, incorporating IT – in this case, multimedia instructional materials – into higher education could improve the quality of learning for students (Alexander, 2001).

Benefits

Several articles (Mbarika, 1999; Mbarika, Sankar & Raju, 2003; Mbarika, Sankar, Raju & Raymond, 2001; Raju & Sankar, 1999; Sankar & Raju, 2002) have evaluated the use of multimedia instructional materials in IS undergraduate classrooms and found the students' responses to be favorable. In using multimedia instructional materials in undergraduate classes, and in our analysis of electronic journals and other students' comments, we have identified the advantages/strengths of multimedia to be as follows:

- Brings theory and practice together in classrooms
- Facilitates the development of higher-order cognitive skills in students
- Provides an informative and fun learning experience
- Encourages active teamwork among students
- Facilitates the development of personal attributes and traits
- Brings excitement of real-world problems into classrooms
- Offers great insight into technology
- Interrelates technical and managerial issues
- Enables and facilitates the development of critical thinking and problem-solving skills.

Limitations

Although this method of instruction has numerous advantages, it is not without its share of limitations/weaknesses. In using multimedia instructional materials in undergraduate classes, we have identified some of the noted limitations/weaknesses to be as follows:

- Requires a heavy investment of energy and planning on the part of the instructor.
- Information may be out of date due to the lengthy development and production cycle of multimedia instructional materials.
- Accreditation agencies may not fully appreciate the uniqueness of such a pedagogy and, thus, may discount its usefulness.

Based on the aforementioned advantages and limitations, it appears evident that the benefits of using multimedia instructional materials outweigh the limitations. Therefore, this next section is aimed at informing faculty members of new ways of teaching using multimedia instructional materials.

STRATEGIES AND INSTRUCTIONS FOR UTILIZATION OF MULTIMEDIA INSTRUCTIONAL MATERIALS

There are few technical case studies that could be directly used in IS classrooms. Our experience in this area suggests that these case studies will be meaningful if they relate to a problem that actually happened in an industry. In support of our belief, Chen (2000) states that using realistic business data facilitates students' problem-solving and decision-making skills, thus better preparing students for what they will face once they leave the classroom. Hence, the development of these case studies should be done in partnership with an industry. We suggest that the technical case studies be peer reviewed and tested in classrooms before they become part of IS curricula.

The case method of teaching requires a heavy investment of instructor energy and planning. It is also a methodology that requires a serious commitment from the student. In light of the various approaches that can be taken to secure student participation, we favor a unique "agreement commitment." A commitment session follows the introductory lecture at which time the instructor and student sign the agreement simultaneously to emphasize the seriousness of the commitment.

We suggest the utilization of two primary tools when using multimedia instructional material – a conventional textbook and a carefully chosen multimedia case study.

Textbook Support

Assuming the materials are being implemented in an introductory IS course, we typically combine the use of traditional textbooks that cover basic introductory concepts in IS and multimedia case studies. The terms multimedia case study and multimedia instructional material are used interchangeably throughout this article. Many introductory textbooks are available for use in IS classrooms; typically, such textbooks do not provide enough in-depth material on the concepts covered. For an introductory IS course, we recommend selecting a textbook whose basic concepts include the following:

- Introduction to IS in organizations
- Hardware and software concepts
- Organizing data and information
- Telecommunications and networks
- Fundamentals of electronic commerce
- Transaction processing systems
- Decision support systems
- Specialized business information systems, such as artificial intelligence, expert systems and virtual reality systems
- Fundamentals of systems analysis and design
- Database management systems concepts
- Information systems security, privacy and ethical issues

Multimedia Case Study Support

We suggest using LITEE multimedia case studies to supplement the theories covered in the textbooks (see www.auburn.edu/research/litee for a list of available case studies). It is of vast importance to choose multimedia case studies that match the topic areas covered in the class. The case studies are packaged in CD-ROM format such that students can use it individually or in teams. The CD-ROMs make it possible for students to see the case study problem visually and, in some cases, hear it spoken audibly by those tasked with making the decision in the real world. The CD-ROMs also include footage of a real person (typically a manager) from the company who explains the issues and leads the students to an assignment. The visual presentation includes factual and live aspects of the case study, such as the problem being investigated, potential alternative solutions to the problem(s) and a request for the students to provide a viable solution. Photos, animations and videos are used to illustrate traditional concepts, thus providing an interactive learning experience for the students. For example, the students can read about hardware concepts from the traditional textbook and then watch video clips, included on the CD-ROM, to gain a better understanding of how the components look, what they do

and how they are designed. The CD-ROM also includes footage on how some of these components can be installed, upgraded or replaced, in addition to providing links to internal and external sources that provide more information about the concepts covered.

The multimedia instructional package also includes a comprehensive instructor's manual in CD-ROM format. The instructor's manual includes video footage showing how the problem was solved in the company. The manual also includes teaching suggestions, PowerPoint presentations and potential exam questions. Both the student version and instructor's manual include several innovative features, such as audio clips, video clips and decision support software.

Using multimedia instructional materials in a classroom requires the work of multiple groups of individuals. The strategies we provide next, though not meant to be exhaustive, are techniques that, in our experiences, have proven effective when using multimedia instructional materials in IS undergraduate classes. A simple analysis of the case study could be performed in one class, whereas a detailed analysis might take 3 to 5 weeks of class time. The Appendix contains samples of lesson plans that may be adapted by those wishing to use multimedia case studies. The lesson plans may be used in the current state or be modified as needed. Due to the large amount of planning that goes into preparing to administer multimedia case studies, we break the lesson plan into three areas – before class, during class and after class.

Before Class

Prior to the initial class session, the instructor should determine the case study to be assigned and provide competency materials to the students. Competency materials relating to the needs of the case study should be developed and shared with the students before they are assigned case studies to analyze. This is different from the traditional case study methodology developed by most business schools. The strategy we propose is essential because of the multi-disciplinary nature of the real-world problems being addressed in the multimedia case studies. It is important to provide background material on the disciplines that have a significant role in the case studies.

Instructors may also use one or more approaches to prepare for class. They may utilize the case teaching notes (TN) that accompany the case as a supplemental resource. The TN provide a summary of the case study, statements of objectives, teaching suggestions and discussion questions with suggested answers. It is also a good idea for instructors to consult with colleagues for additional perspectives. Student preparation for case discussion may involve either writing an analysis that follows an instructor-prescribed format or responding to assigned questions. Small-group discussions preceding the formal class session are encouraged to obtain multiple views and develop student interest in the case specifics.

During Class

Once the class session has begun, the instructor, using the traditional lecture method, may review the competency materials provided prior to the class session. Students are expected to raise questions regarding the readings pertinent to the competency materials. In our experiences, the best approach has been to encourage the students to work in teams whereby they can brainstorm and use other teamwork strategies (covered in a class lecture) to come up with findings/solutions to the case study problem in question. If the class setting is made up of students from multiple disciplines, we suggest building teams that are cross-functional. Teaming exercises and guides might help improve group interaction. The instructor could provide opportunities for different students to lead the team for different case studies, thereby providing opportunity for all students to participate in the discussion. Now that the role of the instructor becomes that of a facilitator at the same time, the instructor has to ensure that students do not steer the class into unrelated topics. The instructor has to encourage students to perform group work. Reference to research material on group work might be helpful to the instructors. The instructor should encourage teams to communicate with each other and the instructor. Tools such as electronic journals, e-mail, discussion boards and chat rooms are very helpful in achieving this objective.

The instructor should emphasize that he/she expects the students to carefully read the technical information in the case studies in order to analyze the problem. Thereafter, the students should be required

to present their findings in class. The presentations are typically made in a competitive manner such that the different teams challenge each other. Students should be encouraged to use multimedia technologies in their presentations. The case analysis part of the session should emphasize participation, led by the instructor, who acts as "facilitator" and "explorer" of the case analysis rather than "master" and "expert." An optional epilogue can be interjected to provide closure to the class session.

After Class

Following the class session, the instructor should evaluate students' contributions either by reviewing the students' written recommendations or by assigning points to their contributions. Separately, the instructor also should evaluate materials and update TN for future sessions. To derive full benefit from the case method, students should exchange their analyses with colleagues and identify how major course concepts applied to the case study.

EVALUATING STUDENT PROGRESS

After using multimedia instructional materials in IS undergraduate classrooms, the next major issue is that of evaluating the students' progress and performance. This evaluation might include the e-journals, presentations and case study write-ups. The instructor should create an evaluation formula to be shared with students prior to the completion of the case study. The clearer the instructor's objectives are to the students, the better the chances are that those expectations will be met. It is critical to establish a mechanism to provide feedback to students about their performance. Evaluation questionnaires similar to the ones used in previous studies (Bradley, Sankar, Clayton & Raju, 2004; Marghitu, Sankar & Raju, 2003; Mbarika, Sankar & Raju, 2003) would provide valuable information on the utility of the selected case studies in the instructor's classrooms. In addition, we recommend that students be requested to submit e-journals, forms with seven to eight questions about the students' thought processes as they progressed through the multimedia instructional material. The e-journals help to document students' progress throughout the course. Since the case studies are performed in teams, each student should submit e-journals in order to evaluate what the students learned individually.

CONCLUSION

This article shares rationales for using multimedia instructional materials and provides instructions on how to use these materials in typical IS undergraduate classrooms. It also includes practical advice for those interested in using multimedia instructional materials, as well as the process of evaluating students in the use of these instructional materials. Research studies show that use of the multimedia instructional materials in IS undergraduate classrooms have the potential to provide enhanced opportunities for active learning. In addition, these instructional materials have been known to stimulate the interest of non-engineering, female and minority students in engineering and technical topics. Thus, using multimedia instructional materials in IS undergraduate classrooms can enhance curriculums and students' experiences.

REFERENCES

Alexander, S. (2001). E-learning developments and Experiences. *Education + Training, 43*(4/5), 240-248.

Barrett, E. (1988). *Text, context, and hypertext*. Cambridge, MA: MIT Press.

Bradley, R.V., Sankar, C.S., Clayton, H., & Raju, P.K. (2004). Using multimedia instructional materials to assess the validity of imposing GPA entrance requirements in colleges of business: An empirical examination. Paper presented at the *15th Annual Information Resources Management Association International Conference*, New Orleans, LA.

Chen, C. (2000). Using realistic business data in teaching business problem solving. *Information Technology, Learning, and Performance Journal, 18*(2), 41-50.

Collier, G.H. (1987). Thoth-II: Hypertext with explicit semantics. Paper presented at the *ACM Conference on Hypertext*, Chapel Hill, NC.

Delany, P., & Gilbert, J.K. (1991). Hypercard stacks for Fielding's Joseph Andrews: Issues of design and content. In P. Delany & G. Landow (Eds.), *Hypertext and literary studies* (pp. 287-298). Cambridge, MA: MIT Press.

Fetterman, R. (1997). *The interactive corporation*. New York: Random House.

Gaytan, J.A., & Slate, J.R. (2002, 2003). Multimedia and the college of business: A literature review. *Journal of Research on Technology in Education, 35*(2), 186-205.

Gorgone, J.T., Gray, P., Feinstein, D., Kasper, G.M., Luftman, J.N., Stohr, E.A., et al. (2000). MSIS 2000 Model curriculum and guidelines for graduate degree programs in information systems. *Communications of the AIS, 3*(1).

Jonassen, D.H. (1989). *Hypertext/Hypermedia*. Englewood Cliffs: Education Technology Publications.

Marghitu, D., Sankar, C.S., & Raju, P.K. (2003). Integrating a real life engineering case study into the syllabus of an undergraduate network programming using HTML and JAVA course. *Journal of SMET Education, 4*(1/2), 37-42.

Mbarika, V. (1999). An experimental research on accessing and using information from written vs. multimedia systems. Paper presented at the *Fifth Americas Conference on Information Systems*, Milwaukee, WI.

Mbarika, V., Sankar, C.S., & Raju, P.K. (2003). Identification of factors that lead to perceived learning improvements for female students. *IEEE Transactions on Education, 46*(1), 26-36.

Mbarika, V., Sankar, C.S., Raju, P.K., & Raymond, J. (2001). Importance of learning-driven constructs on perceived skill development when using multimedia instructional materials. *Journal of Educational Technology Systems, 29*(1), 67-87.

Nielsen, J. (1995). *Multimedia and hypertext: The Internet and beyond*. Boston: AP Professional.

Nunamaker, J.F., Jr., Couger, J.D., & Davis, G.B. (1982). Information systems curriculum recommendations for the 80s: Undergraduate and graduate programs. *Communications of the ACM, 25*(11), 781-805.

Oliver, R., & Omari, A. (1999). Investigating implementation strategies for WWW-based learning environments. *International Journal of Instructional Media, 25*(2), 121-136.

Raju, P.K., & Sankar, C.S. (1999). Teaching real-world issues through case studies. *Journal of Engineering Education, 88*(4), 501-508.

Sankar, C.S., & Raju, P.K. (2002). Bringing real-world issues into classrooms: A multimedia case study approach. *Communications of the AIS, 8*(2), 189-199.

APPENDIX OF SAMPLE LESSON PLANS FOR THE USE OF MULTIMEDIA INSTRUCTIONAL MATERIALS

5-Week Plan

Week 1:
Introduction, Team Building Exercises

Week 2:
Divide students into teams
Lecture: Assign Case Study
Lab: Case Study student work session

Week 3:
Lecture: Technical & Business Issues (from traditional textbook)
Lab: Case Study student work session

Week 4:
Lecture: Technical & Business Materials (from traditional textbook)
Lab: Case Study Presentations

Week 5:
Lecture: What Happened? Feedback session on case study, and e-journals

1-Week Plan (based on two class meetings)

Day 1:
Introduction, Divide students into teams, Assign Case Study, Teach competency materials (from traditional textbook)

Day 2:
Case Study Presentations
Day 2:
Last 15 minutes: Lecture: What Happened? Feedback session on case study

1-Day Plan

Session 1:
Introduction, Divide students into teams, Assign Case Study, Teach competency materials (from traditional textbook)
Session 2:
Case Study Presentations
Last 15 minutes: Lecture: What Happened? Feedback session on case study

KEY TERMS

Bookshelf: The combination of text, graphics, audio and video with tools that allows the user to navigate, interact, create and communicate the content or his or her own ideas, but lacks the links to connect the information.

E-Journal: Electronic form with a series of questions (e.g., seven to eight) that help to document students' progress throughout the course, pertaining to the students' thought processes as they progress through the multimedia instructional material.

LITEE (Laboratory for Innovative Technology and Engineering Education): National Science Foundation-sponsored research group at Auburn University that develops award-winning multimedia instructional materials that bring theory, practice and design together for the purpose of bringing real-world issues into engineering and business classrooms.

Mixed Media: The combination of text, graphics, audio and video with links and tools that allows the user to navigate, create and communicate the content or his or her own ideas, but lacks the component that provides interactivity.

Movie: The combination of text, graphics, audio and video with links and tools that allows the user to interact, create and communicate the content or his or her own ideas, but lacks navigational tools that would allow the user to choose his or her course of action.

Multimedia: The use of a computer to present and combine text, graphics, audio and video with links and tools that allows the user to navigate, interact, create and communicate the content or his or her own ideas. Multimedia is sometimes referred to as "Interactive Media."

Television: The combination of text, graphics, audio and video with links and tools that allows the user to navigate and interact, but lacks the means to provide users the opportunity to create contribute their own ideas.

ENDNOTES

[1] The materials reported in this article are based partially upon work supported by the National Science Foundation under Grant Numbers 9950514, 0089036 and 0527328. Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[2] LITEE, Laboratory for Innovative Technology and Engineering Education, is an NSF-sponsored project conducted at Auburn University that creates award-winning multimedia instructional materials. Its instructional materials are reported as being helpful in facilitating the improvement of students' higher-order cognitive skills. LITEE multimedia instructional materials cover concepts ranging from strategic management of IT and decision support to financial management of IT investments. Information about multimedia instructional materials available from LITEE can be found at www.auburn.edu/research/litee.

Multimedia Interactivity on the Internet

Omar El-Gayar
Dakota State University, USA

Kuanchin Chen
Western Michigan University, USA

Kanchana Tandekar
Dakota State University, USA

INTRODUCTION

With the interactive capabilities on the Internet, business activities such as product display, order placing and payment are given a new facelift (Liu & Shrum, 2002). Consumer experience is also enhanced in an interactive environment (Haseman, Nuipolatoglu & Ramamurthy, 2002). A higher level of interactivity increases the perceived telepresence and the user's attitude towards a Web site (Coyle & Thorson, 2001). When it comes to learning, a higher level of interactivity improves learning and learner satisfaction (Liu & Schrum, 2002). While interactivity does not necessarily enable enhanced gain in user learning, it positively influences learners' attitudes (Haseman et al., 2002). Interactivity has been shown to engage users in multimedia systems (Dysart, 1998) to encourage revisits to a Web site (Dholakia et al., 2000), to increase satisfaction toward such systems (Rafaeli &Sudweeks, 1997), to enhance the visibility (as measured in number of referrals or backward links) of Web sites (Chen & Sockel, 2001) and to increase acceptance (Coupey, 1996).

BACKGROUND

According to the Merriam Webster dictionary, "interactivity" refers to 1) being mutually or reciprocally active, or 2) allowing two-way electronic communications (as between a person and a computer). However, within the scientific community, there is little consensus of what interactivity is, and the concept often means different things to different people (Dholakia, Zhao, Dholakia & Fortin, 2000; McMillan & Hwang, 2002). McMillan and Hwang (2002) suggest that interactivity can be conceptualized as a process, a set of features and user perception. Interactivity as a process focuses on activities such as interchange and responsiveness. Interactive features are made possible through the characteristics of multimedia systems. However, the most important aspect of interactivity lies in user perception of or experience with interactive features. Such an experience may very likely be a strong basis for future use intention.

Interactivity is considered a process-related construct, where communication messages in a sequence relate to each other (Rafaeli & Sudweeks, 1997). Ha and James (1998, p. 461) defined interactivity as "the extent to which the communicator and the audience respond to, or are willing to facilitate, each other's communication needs." Interactions between humans via media are also called mediated human interactions or computer-mediated communication (Heeter, 2000). Early studies tend to consider interactivity as a single construct, where multimedia systems vary in degrees of interactivity. Recent studies suggest that interactivity is a multidimensional construct.

As research continues to uncover the dynamic capabilities of multimedia systems, the definition of interactivity evolves to include aspects of hardware/software, processes during which the interactive features are used and user experience with interactive systems. Dholakia et al. (2000) suggest the following six interactivity dimensions: 1) user control, 2) responsiveness, 3) real-time interactions, 4) connectedness, 5) personalization/customization, and 6) playfulness. Similarly, Ha and James (1998)

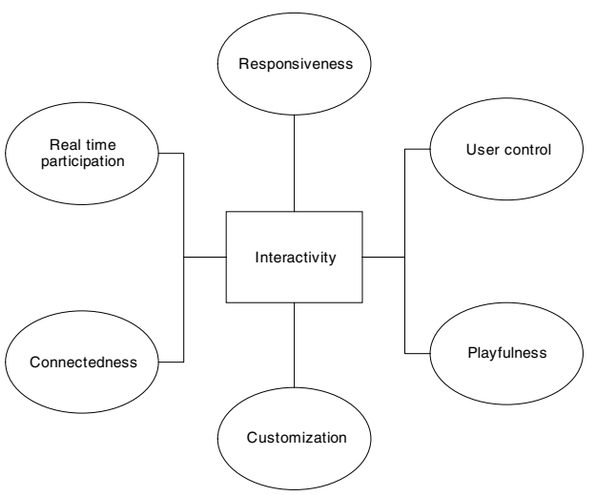
Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

suggest five interactivity dimensions: 1) playfulness, 2) choice, 3) connectedness, 4) information collection, and 5) reciprocal communication.

Within the context of multimedia systems, we view interactivity as a multidimensional concept referring to the nature of person-machine interaction, where the machine refers to a multimedia system. Figure 1 presents a conceptual framework, including interactivity dimensions defined as follows:

- **User control:** The extent to which an individual can choose the timing, content and sequence of communication with the system.
- **Responsiveness:** The relatedness of a response to earlier messages (Rafaeli & Sudweeks, 1997).
- **Real-time participation:** The speed with which communication takes place. This can range from instant communication (synchronous) to delayed response communication (asynchronous).
- **Connectedness:** The degree to which a user feels connected to the outside world through the multimedia system (Ha & James, 1998).
- **Personalization/Customization:** The degree to which information is tailored to meet the needs of individual users. For example, interactive multimedia learning systems must be able to accommodate different learning styles and capabilities.

Figure 1. Interactivity as a multidimensional concept



- **Playfulness:** The entertainment value of the system; that is, entertainment value provided by interactive games or systems with entertaining features.

TECHNOLOGIES AND PRACTICES

The ubiquity of multimedia interactivity in general and on the Internet in particular is realized through the exponential growth in information technology. Specifically, the growth in computational power enabling ever-increasingly multimedia features coupled with advances in communication technologies and the Internet are pushing the interactivity frontier. Such technologies include, but are not limited to, a range of technologies, from the basic point and click to highly complex multimedia systems.

In practice, and in their quest for interactivity, companies and organizations have resorted to a variety of techniques to encourage interactions in their systems. Table 1 provides a framework to map important multimedia/Web features from the existing literature to the six interactivity dimensions discussed in Figure 1. The goal of this framework is to offer practitioners a basis to evaluate interactivity in their multimedia systems. For example, a Web site designer may want to compare his or her design with popular Web sites in the same industry to measure if they offer a similar level of interactivity. Two important issues concerning the comparison include what interactive features are recommended for comparison and how to quantify interactivity features for comparison. The framework in Table 1 serves to answer the first question. One way to answer the second question involves simply counting the number of interactivity features in each of the interactivity dimensions. This counting technique is referred to as the interactivity index (II) and is frequently used by researchers to quantify interactivity. The quantified results, if measured consistently, can be used for longitudinal or cross-industry comparisons. Additionally, interactivity is examined with other constructs. Readers interested in empirical results focusing on the relationship between interactivity dimensions and other constructs are referred to the cited references, such as Ha and James (1998), Dholakia et al. (2000); Chen and Sockel (2001); McMillan and Hwang (2002); Burgoon, Bonito,

Ramirez, Dunbar, Kam and Fischer (2002); and Chen and Yen (2004).

CURRENT RESEARCH

Interactivity is an active area of research that spans a number of research fields, including computer science, human computer interaction (HCI), information systems, education, marketing, advertisement and communication. A comprehensive review of the literature is beyond the scope of this article. Instead, we focus our attention on current research effort as it pertains to multimedia interactivity on the Internet, with a particular emphasis on education, advertisement and marketing.

Current research on multimedia interactivity predominantly focuses on conceptual issues related to the definition and measurement of interactivity, evaluation of interactive multimedia systems, design issues and applications of interactive multimedia systems. Regarding conceptual issues, Kirch (1997) questions the decision cycle model, which is the received theory in human computer interaction, and discusses additional ways of interacting with multimedia systems; while Ohl (2001) questions the adequacy of current definitions of interactivity in the context of educational systems.

Haseman, Polatoglu and Ramamurthy (2002) found that interactivity leads to favorable attitude formation but not so much to improved learning outcomes. There had been no evidence to prove that interactivity influences user achievement. Liu and Shrum (2002) propose that higher levels of interactivity create a cognitively involving experience and can enhance user satisfaction and learning.

Concerning design considerations, Robinson (2004) identifies interactivity as one of eight principles for the design of multimedia material. Examples of case studies and applications reported in the literature include Abidin and Razak's (2003) presentation of Malay folklore using interactive multimedia. Table 2 lists research contributions pertaining primarily to multimedia interactivity.

Internet interactivity has also attracted interest in areas such as the measurement of interactivity, evaluation of the effectiveness of interactivity and design considerations for Internet-interactive Web sites. For example, Paul (2001) analyzed the content of 64 disaster relief Web sites and found that most sites had a moderate level of interactivity but were not very responsive to their users. A study conducted by Ha and James (1998) attempted to deconstruct the meaning of interactivity, and then reported the results of a content analysis that exam-

Table 1. A framework of mapping multimedia/Web features to interactivity dimensions

Interactivity dimensions	Multimedia/Web features	
User control	• Alternative options for site navigation • Linear interactivity, where the user is able to move (forward or backwards) through a sequence of contents	• Object interactivity (proactive inquiry) where objects (buttons, people or things) are activated by using a pointing device.
Responsiveness	• Context-sensitive help • Search engine within the site	• Dynamic Q&A (questions and responses adapt to user inputs)
Real-time participation	• Chat rooms • Video conferencing	• E-mail • Toll-free number
Connectedness	• Video clips • Site tour	• Audio clips • Product demonstration
Personalization/ Customization	• Site customization • Bilingual site design	• Customization to accommodate browser differences
Playfulness	• Games • Software downloads • Visual simulation	• Online Q&A • Browser plug-ins (e.g., flash, macromedia, etc.)

Table 2. Current research focusing primarily on multimedia interactivity

Research focus	Conceptual	Evaluation	Design	Application
Research work	Ohl (2001), Massey (2000)	Karayanni et al. (2003), Haseman et al. (2002), Liu and Shrum (2002), Ellis (2001), Moreno (2001), Mayer (2001)	Robinson (2004), Zhang et al. (2003), Trindade et al. (2002)	Adibin et al. (2003), Hou et al. (2002), Paustian (2001)

ined the interactivity levels of business Web sites. Their findings suggest that five interactivity dimensions are possible, with the reciprocal communication dimension being the most popular dimension. In an effort to explore the relationship between Ha and James' (1998) interactivity dimensions and the quality of Web sites, Chen and Yen (2004) suggested that reciprocal communication, connectedness and playfulness are the most salient dimensions of interactivity that influence design quality. Moreover, Lin and Jeffres (2001) performed a content analysis of 422 Web sites associated with local newspapers, radio stations and television stations in 25 of the largest metro markets in the United States. Results show that each medium has a relatively distinctive content emphasis, while each attempts to utilize its Web site to maximize institutional goals.

According to Burgoon et al. (2002), computer mediated communication may even be better than non-mediated or face-to-face interaction, even though face-to-face is considered easier. The study also points out that distal communication, mediation and loss of non-verbal cues do not necessarily result in worse decision quality or influence, but may, in fact, enhance performance in some cases.

Addressing design considerations, McMillan (2000) identified 13 desirable features that an interactive Web site should possess in order to be interactive. These features include: e-mail links, hyperlinks, registration forms, survey forms, chat rooms, bulletin boards, search engines, games, banners, pop-up ads, frames and so forth. High levels of vividness help create more enduring attitudes (Coyle & Thorson, 2001). A study by Bucy, Lang, Potter and Grabe (1999) found presence of advertising in more than half of the Web pages sampled. It also suggests a possible relationship between Web site traffic and the amount of asynchronous interactive elements like text links, picture links, e-mail links, survey forms and so forth. Features most commonly used on the surveyed Web sites were frames, logos and a white background color.

FUTURE TRENDS

Long-term impacts of interactivity should be studied on learning, attitudes and user outcomes. To study learning behavior of students, their knowledge should be tested twice; once at first and then after a few days or weeks for absorption/retention (Haseman et al., 2002). Coyle and Thorson (2001, p. 76) suggested to "focus on additional validation of how new media can approximate a more real experience than traditional media." One way to do this would be to replicate previous findings dealing with direct or indirect experience. Will more interactive and more vivid systems provide more direct experience than less interactive, less vivid systems? Also, future research should focus on testing specific tools to understand how their interactivity characteristics improve or degrade the quality of user tasks at hand.

The current literature appears to lack consensus on the dimensionality of interactivity. Inconsistent labeling or defining the scope of interactivity dimensions exists in several studies; for example, playfulness and connectedness appear to be included in both Dholakia et al. (2000) and Ha and James (1998), but Dholakia et al.'s personalization/customization dimension was embedded in Ha and James' choice dimension. Furthermore, much of interactivity research employed only qualitative assessment of interactivity dimensions (such as Heeter, 2000), suggesting future avenues for empirical validations and perhaps further refinement.

Despite disagreements in interactivity dimensions, user interactivity needs may vary across time, user characteristics, use contexts and peer influ-

ence. A suggestion for further research is to take into account the factors that drive or influence interactivity needs in different use contexts. Another suggestion is to study whether user perception depends on the emotional, mental and physical state of people; that is, their personality and to what extent or degree it depends on these characteristics and how these can be altered to improve the overall user perception.

CONCLUSION

Multimedia interactivity on the Internet – while considered as "hype" by some – is here to stay. Recent technological advancements in hardware, software and networks have enabled the development of highly interactive multimedia systems. Studying interactivity and its effects on target users certainly impact business values. Research pertaining to interactivity spans a number of disciplines, including computer science, information science, education, communication, marketing and advertisement. Such research addressed a variety of issues, ranging from attempting to define and quantify interactivity to evaluating interactive multimedia systems in various application domains, to designing such systems. Nevertheless, a number of research issues warrant further consideration, particularly as it pertains to quantifying and evaluating interactive multimedia systems.

In effect, the critical issues discussed in this chapter offer many implications to businesses, governments and educational institutions. With regard to businesses, multimedia interactive systems will continue to play a major role in marketing and advertisement. Interactive virtual real estate tours are already impacting the real estate industries. Interactive multimedia instruction is changing the way companies and universities alike provide educational services to their constituents. From physics and engineering to biology and history, interactive multimedia systems are re-shaping education.

REFERENCES

Abidin, M.I.Z., & Razak, A.A. (2003). Malay digital folklore: Using multimedia to educate children through storytelling. *Information Technology in Childhood Education Annual*, (1), 29-44.

Bucy, E.P., Lang, A., Potter, R.F., & Grabe, M.E. (1999). Formal features of cyberspace: Relationship between Web page complexity and site traffic. *Journal of the American Society for Information Science, 50*(13), 1246-1256.

Burgoon, J.K., Bonito, J.A., Ramirez, A., Dunbar, N.E., Kam, K., & Fischer, J. (2002). Testing the interactivity principle: Effects of mediation, propinquity, and verbal and nonverbal modalities in interpersonal interaction. *Journal of Communication, 52*(3), 657-677.

Chen, K., & Sockel, H. (2001, August 3-5). Enhancing visibility of business Web sites: A study of cyber-interactivity. *Proceedings of Americas Conference on Information Systems*, (pp. 547-552).

Chen, K., & Yen, D.C. (2004). Improving the quality of online presence through interactivity. *Information & Management,* forthcoming.

Coupey, E. (1996). Advertising in an interactive environment: A research agenda. In D.W. Schumann & E. Thorson (Eds.), *Advertising and the World Wide Web* (pp. 197-215). Mahwah, NJ: Lawrence Erlbaum Associates.

Coyle, J.R., & Thorson, E. (2001). The effects of progressive levels of interactivity and vividness in Web marketing sites. *Journal of Advertising*, *30*(3), 65-77.

Dholakia, R.R., Zhao, M., Dholakia, N., & Fortin, D.R. (2000). Interactivity and revisits to Web sites: A theoretical framework. Research institute for telecommunications and marketing. Retrieved from *http://ritim.cba.uri.edu/wp2001/wpdone3/Interactivity.pdf*

Dysart, J. (1998). Interactivity: The Web's new standard. *NetWorker: The Craft of Network Computing*, *2*(5), 30-37.

Ellis, T.J. (2001). Multimedia enhanced educational products as a tool to promote critical thinking in adult students. *Journal of Educational Multimedia and Hypermedia*, *10*(2), 107-124.

Ha, L. (2002, April 5-8). Making viewers happy while making money for the networks: A compari-

son of the usability, enhanced TV and TV commerce features between broadcast and cable network Web sites. *Broadcast Education Association Annual Conference*, Las Vegas, Nevada.

Ha, L., & James, E.L. (1998). Interactivity reexamined: A baseline analysis of early business Web sites. *Journal of Broadcasting & Electronic Media*, *42*(4), 457-474.

Haseman, W.D., Nuipolatoglu, V., & Ramamurthy, K. (2002). An empirical investigation of the influences of the degree of interactivity on user-outcomes in a multimedia environment. *Information Resources Management Journal*, *15*(2), 31-41.

Heeter, C. (2000). Interactivity in the context of designed experiences. *Journal of Interactive Advertising*, *1*(1). Available at *www.jiad.org/vol1/no1/heeter/index.html*

Hou, T., Yang, C., & Chen, K. (2002). Optimizing controllability of an interactive videoconferencing system with Web-based control interfaces. *The Journal of Systems and Software*, *62*(2), 97-109.

Karayanni, D.A., & Baltas, G.A. (2003). Web site characteristics and business performance: Some evidence from international business-to-business organizations. *Marketing Intelligence & Planning*, *21*(2), 105-114.

Lin, C.A., & Jeffres, L.W. (2001). Comparing distinctions and similarities across Web sites of newspapers, radio stations, and television stations. *Journalism and Mass Communication Quarterly*, *78*(3), 555-573.

Liu, Y., & Shrum, L.J. (2002). What is interactivity and is it always such a good thing? Implications of definition, person, and situation for the influence of interactivity on advertising effectiveness. *Journal of Advertising*, *31*(4), 53-64.

Massey, B.L. (2000). Market-based predictors of interactivity at Southeast Asian online newspapers. *Internet Research*, *10*(3), 227-237.

Mayer, R.E., & Chandler, P. (2001). When learning is just a click away: does simple user interaction foster deeper understanding of multimedia messages? *Journal of Educational Psychology*, *93*(2), 390-397.

McMillan, S.J. (2000). Interactivity is in the eye of the beholder: Function, perception, involvement, and attitude toward the Web site. In M.A. Shaver (Ed.), *Proceedings of the American Academy of Advertising* (pp. 71-78). East Lansing: Michigan State University.

McMillan, S. J., & Hwang, J. (2002). Measures of perceived interactivity: An exploration of the role of direction of communication, user control, and time in shaping perceptions of interactivity. *Journal of Advertising*, *31*(3), 29-42.

Moreno, R., Mayer, R.E., Spires, H., & Lester, J. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, *19*(2), 177-213.

Ohl, T.M. (2001). An interaction-centric learning model. *Journal of Educational Multimedia and Hypermedia*, *10*(4), 311-332.

Paul, M.J. (2001). Interactive disaster communication on the Internet: A content analysis of 64 disaster relief. *Journalism and Mass Communication Quarterly*, *78*(4), 739-753.

Paustian, C. (2001). Better products through virtual customers. *MIT Sloan Management Review*, *42*(3), 14.

Rafaeli, S., & Sudweeks, F. (1997). Networked interactivity. *Journal of Computer-Mediated Communication*, *2*(4). Available at *www.ascusc.org/jcmc/vol2/issue4/rafaeli.sudweeks.html*

Robinson, W.R. (2004). Cognitive theory and the design of multimedia instruction. *Journal of Chemical Education*, *81*(1), 10.

Trindade, J., Fiolhais, C., & Almeida, L. (2002). Science learning in virtual environments: a descriptive study. *British Journal of Educational Technology*, *33*(4), 471-488.

Zhang, D., & Zhou, L. (2003). Enhancing e-learning with interactive multimedia. *Information Resources Management Journal*, *16*(4), 1-14.

KEY TERMS

Computer-Mediated Communication (CMC): Refers to the communication that takes place between two entities through a computer, as opposed to face-to-face interaction that takes place between two persons present at the same time in the same place. The two communicating entities in CMC may or may not be present simultaneously.

Machine Interactivity: Interactivity resulted from human-to-machine or machine-to-machine communications. Typically, the later form is of less interest to most human-computer studies.

Reach: To get users to visit a Web site for the first time. It can be measured in terms of unique visitors to a Web site.

Reciprocal Communication: Communication that involves two or more (human or non-human) participants. The direction of communication may be two way or more. However, this type of communication does not necessarily suggest that participants communicate in any preset order.

Stickiness: To make people stay at a particular Web site. It can be measured by time spent by the user per visit.

Synchronicity: It refers to the spontaneity of feedback received by a user in the communication process. The faster the received response, the more synchronous is the communication.

Telepresence: Defined as the feeling of being fully present at a remote location from one's own physical location. Telepresence creates a virtual or simulated environment of the real experience.

Two-Way Communication: Communication involving two participants; either both of the participants can be humans or it could be a human-machine interaction. It does not necessarily take into account previous messages.

Multimedia Proxy Cache Architectures

Mouna Kacimi
University of Bourgogne, France

Richard Chbeir
University of Bourgogne, France

Kokou Yetongnon
University of Bourgogne, France

INTRODUCTION

The Web has become a significant source of various types of data, which require large volumes of disk space and new indexing and retrieval methods. To reduce network load and improve user response delays, various traditional proxy-caching schemes have been proposed (Abonamah, Al-Rawi, & Minhaz, 2003; Armon & Levy, 2003; Chankhunthod, Danzig, Neerdaels, Schwartz, & Worrell, 1996; Chu, Rao, & Zhang, 2000; Fan, Cao, Almeida, & Broder, 2000; Francis, Jamin, Jin, Jin, Raz, Shavitt, & Zhang, 2001; Paul & Fei, 2001; Povey & Harrison, 1997; Squid Web Proxy Cache, 2004; Wang, Sen, Adler, & Towsley, 2002). A proxy is a server that sits between the client and the real server. It intercepts all queries sent to the real server to see if it can fulfill them itself. If not, it forwards the query to the real server. A cache is a disk space used to store the documents loaded from the server for future use. A proxy cache is a proxy having a cache. The characteristics of traditional caching techniques are threefold. First, they regard each cached object as having no dividable data, which must be recovered and stored in their entirety. As multimedia objects like videos are usually too large to be cached in their entirety, the traditional caching architectures cannot be efficient for this kind of object. Second, they do not take into account the data size to manage the space storage. Third, they do not consider in their caching-system design the timing constraints that need moving objects.

The size is the main difference between multimedia and textual data. For instance, if we have a 2-hour-long MPEG movie, we need around 1.5 Gb of disk space. Given a finite storage space, only a few streams could be stored in the cache, thus, it would decrease the efficiency of the caching system. As the traditional techniques are not efficient for media objects, some multimedia caching schemes have been proposed (Guo, Buddhikot, & Chae, 2001; Hofmann, Eugene Ng, Guo, Paul, & Zhang, 1999; Jannotti, Gifford, Johnson, Kaashoek, & O'Toole, 2001; Kangasharju, Hartanto, Reisslein, & Ross Keith, 2001; Rejaie, Handley, Yu, & Estrin, 1999; Rejaie & Kangasharju, 2001). Two main categories of multimedia caching solutions can be distinguished.

- The first category is storage oriented; it defines new storage mechanisms appropriated to data types in order to reduce the required storage space.
- The second category is object-transmission oriented; it gives new transmission techniques providing large cooperation between proxies to transfer requested objects and reduce bandwidth consumption.

In this article, we briefly present a dynamic multimedia proxy scheme based on defining the profile that is used to match the capacities of the proxies to the user demands. Users with the same profile can easily and quickly retrieve the corresponding documents from one or several proxies having the same profile. A key feature of our approach is the routing profile table (RPT). It is an extension of the traditional network routing table used to provide a global network view to the proxies. Another important feature of the approach is the ability to dynamically adapt to evolving network connectivity: When a proxy is connected to (or disconnected from) a group, we define different schemes for updating the routing

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

profile table and the contents of the corresponding caches. Furthermore, our approach stores data and/or metadata in function of storage capacities of each proxy (For instance, if storage capacities are minimum, only textual and metadata are cached.).

The remainder of the article is organized as follows. The next section gives a snapshot of current proxy-caching techniques provided in the literature. The section after that presents our approach, and then we finally conclude the article and give our future work.

BACKGROUND

Two main categories of traditional (or textual-oriented) caching solutions can be distinguished in the literature: hierarchical caching and distributed caching. In a hierarchical caching architecture (Chankhunthod et al., 1996), the caches are organized in several levels. The bottom level contains client caches and the intermediate levels are devoted to proxies and their associated caches. When a query is not satisfied by the local cache, it is redirected to the upper level until there is a hit at a cache. If the requested document is not found in any cache, it is submitted directly to its origin server. The returned document is sent down the cache hierarchy to the initial client cache and a copy is left on all intermediate caches to which the initial user requests were submitted. Hierarchical caching has many advantages; it avoids the redundant retrieval of documents from data servers, reduces network bandwidth demands, and allows the distribution of document accesses over a cache's hierarchy. Despite its advantages, hierarchical caching exhibits several drawbacks. The high-level caches, particularly the root cache, are bottlenecks that can significantly degrade the performance of the system. The failure of a cache can affect the system fault tolerance. Several copies of the same document are stored at different cache levels, which is very storage expensive and restrictive, especially when treating multimedia data. Moreover, there is a lack of direct links between sibling caches of the same level.

The distributed caching approach reduces hierarchical links between caches. Several distributed caching approaches have been proposed to address one or more problems associated with hierarchical caching (Armon & Levy, 2003; Fan et al., 2000; Povey & Harrison, 1997). In Povey and Harrison, the authors propose an extension of hierarchical caching where documents are stored on leaf caches only. The upper level caches are used to index the contents of the lower level caches. When a query cannot be satisfied by the local cache, it is sent to the parent cache that indicates the location of the required documents. In Fan et al., the authors propose a scalable distributed cache approach, called summary cache, in which each proxy stores a summary of its cached-documents directory on every other proxy. When a requested document is not found in the local cache, the proxy checks the summaries in order to determine relevant proxies to which it sends the request to fetch the required documents. Two major problems restrain the scalability of the summary-cache approach. The first problem is the frequency of summary updates, which can significantly increase interproxy traffic and bandwidth usage. The second problem is related to the storage of the summaries, especially when the number of cooperating proxies is important. Armon and Levy investigate the cache satellite distribution system, which comprises P proxy caches and a central station. The central station periodically receives from the proxy caches reports containing information about user requests. The central station uses this information to foresee what documents could be required by other proxy caches in the near future. It selects a collection of popular Web documents and broadcasts the selected documents via satellite to all or some of the participating proxies. There are two important advantages of this proposal: (a) It anticipates user requests, and (b) it allows collaboration between proxies independent from the geographical distance of a satellite. However, the central station used leads to a weak fault tolerance.

As for textual-oriented caching solutions, two main categories of multimedia caching solutions can be distinguished: storage oriented and object-transmission oriented. In the first category, basic techniques (Acharya & Smith, 2000; Guo et al., 2001) consist of dividing each media data into small segments and distributing them among different caches. The segment distribution helps to have a virtual storage space; that is, the user can store data even if the local cache does not have sufficient free space. In this manner, a cooperative caching schema is defined. When a client requests a multimedia object, the

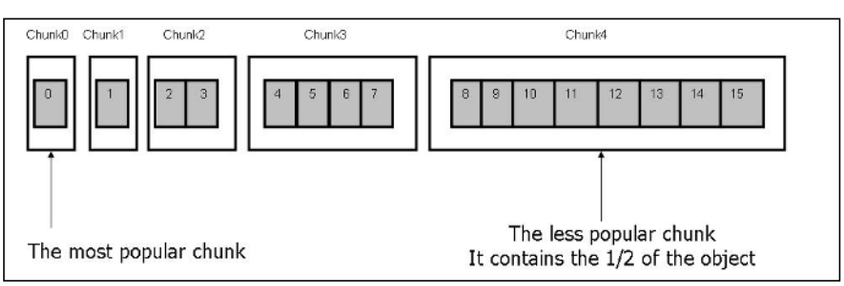
corresponding segments are recovered from several caches. This cooperation allows reducing latency if the cooperative caches belong to the same neighborhood. Otherwise, this approach can introduce additional delays. In MiddleMan (Acharya & Smith), only one copy of each segment is stored, which is very restrictive whenever one of the caches containing a requested segment is down. To resolve the problem of fault tolerance, RCache (Guo et al.) suggests storing a variable number of each segment in function of its global and local popularity. However, it does not address the storage-space consumption problem.

To improve the use of the space storage, PCMS (Segment-Based Proxy Caching of Multimedia Streams) (Wu, Yu, & Wolf, 2001) proposes an interesting cache-management technique based on the importance of starting segments. Only the first segments are stored and the remaining ones are fetched from the origin server when requested. This technique reduces both the storage-space consumption and the user response delays because the start-up latency depends on the starting segments. Therefore, the user does not need to wait for a long time to play the media clip. The popularity metric is not used only to define segments that must be stored, but also to define which ones must be removed when the storage space is exhausted. PCMS presents a new segmentation method based on the popularity in order to improve the existing replacement policies. It defines a logic unit called a chunk. The chunk is a set of consecutive segments, and the media is a set of chunks. The smaller unit of storage and transfer is the segment, and the smaller unit of replacement is the chunk. This approach focuses on the importance of the starting segments. Hence, it defines the size of the chunk in function of its distance from the beginning of the media. It means that the closer chunk to the beginning is the smaller one. Using this sized segmentation, the cache manager can discard half of the cached objects in a single action since the chunks having the lower popularity are the large ones. In contrast, using the simple segmentation, the cache manager can do the same action taking more time.

In the second category of multimedia caching, several approaches have been proposed (Jannotti et al., 2001; Kangasharju et al., 2001; Rejaie & Kangasharju, 2001) in order to adapt the transmission rate in function of the available bandwidth. These approaches, called adaptive, consist of storing video as encoded layers. Each video is encoded using a base layer with one or more enhancement layers. The base layer contains basic information while enhancement layers contain the complementary data. A particular enhancement layer can only be decoded if all lower layers are available. Given the presence of layered video in the origin server, the problem is to determine which videos and layers should be cached.

To define the transmission strategy of video layers, Kangasharju et al. assign a quality to each stored video. The quality depends on the number of layers. It means that video streams with n quality correlate to n layers available in the cache. Using the quality, the video-layers transmission is done as follows. The user sends a request with j-quality video streams to the appropriate proxy cache. If all the requested layers are stored in the cache, the user recovers them. Otherwise, a connection is established with the origin server if there is sufficient bandwidth to retrieve the missing layers. In the case where no sufficient bandwidth is available, the request is blocked and the service provider tries to offer a lower quality stream of the requested object. The main advantage of encoded-layers video is adapting the transmission rate in function of the available

Figure 1. Chunks



bandwidth. This adaptation maximizes the storage efficiency while minimizing the latency time and the load on the network.

The existing multimedia caching techniques use a separate unicast stream for each request. Thus, the server load and the latency increase with the number of receivers and the network congestion, respectively. Hofmann et al. (1999) give a solution to this problem by providing a dynamic caching approach. The difference between classical caching, called static caching and dynamic caching, is the data deliverance strategy. In static caching, two playback requests require two separate data streams. In dynamic caching, the same data stream is shared between two requests.

One of the drawbacks of current caching techniques is that they are too restrictive. They only provide a static architecture not adaptable to network evolution (new materials, fault tolerance, etc.) and user-demands evolution (new users, new profiles, etc.). For instance, the disconnection or connection of a proxy (even if it was the root one) on the network should be managed dynamically in function of the network traffic and server capacities. Furthermore, when defining Web caching schemes for multimedia applications, major issues should be addressed, which are the optimization of the storage capacities and the improvement of information retrieval.

DYNAMIC MULTIMEDIA PROXY

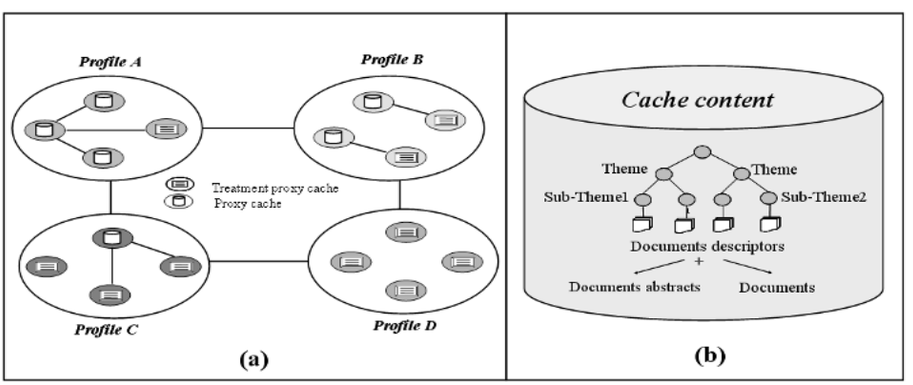
To improve multimedia data-retrieval relevance, anticipate end-user queries, and enhance network performances, our approach consists of the following.

- Organizing machines by profile groups (Figure 2a). In our approach, the profile describes user interests and preferences in terms of themes (sports, cinema, news, etc.). Therefore, users and caches having the same interests can easily and quickly exchange related documents.
- Optimizing storage by selecting only profile-related documents on caches
- Considering the machines' capacities in terms of storage, treatment, and communication
- Providing a quick and pertinent answer
- Maintaining a global vision on each proxy of the network in order to optimize traffic and quickly forward queries to appropriate proxies
- Stringing documents in an indexing tree (Figure 2b) to improve their retrieval

Proxy Types

In LAN (local-area network) or WAN (wide-area network), machines have different capacities in terms of treatment, storage, and communication. For this reason, we have defined two types of proxies: proxy cache and treatment proxy cache. A proxy cache is identified as a machine that has low capacities for running parallel or specialized treatments, and for managing communications between the proxies. Its main task is to store multimedia data and other metadata such as local indexes, neighbor indexes, routing profile tables, and document descriptions. A treatment proxy cache is identified as a powerful

Figure 2. Profile groups and proxy types



machine with high capacities for storage, treatment, and/or communication. In addition to its storage capacity, its main task is to manage a set of cache proxies of the same profile and to maintain the index of their content.

Index and Routing Profile Table

In our approach, each proxy must recognize its cache content and its environment. The processes of recognition and multimedia document retrieval are based on a set of indexes allowing the grouping of the proxies. There are three types of indexes: local indexes, neighbor indexes, and routing profile tables. The local index is used to process user queries and to fetch documents in the local cache. The neighbor index allows forwarding queries toward treatment proxies in function of the profile and the connection time in order to retrieve the corresponding documents. The routing profile table is an extension of the classical routing table, which gives a global view of the network. It contains a list of couples (treatment proxy, profile) that allows a proxy cache to choose, in function of a profile, the treatment proxy to which queries will be sent. The RPT is located on each proxy (cache and treatment) and is also used for connection and disconnection purposes.

CONCLUSION AND FUTURE TRENDS

We have presented in this article an overview of the textual-oriented and multimedia-oriented caching schemes. Textual-oriented approaches are divided into two categories: hierarchical caching and distributed caching. Similarly, two categories can be distinguished in the multimedia approaches: storage-oriented and transmission-object-oriented techniques. We presented each approach, addressing its advantages and limits. We also presented our approach of a dynamic multimedia proxy. It evolves with the network usage and the user demands. We believe that our proposition is able to optimize storage capacities and improve information-retrieval relevance.

Our future directions are various. First, real-life case studies are needed to deploy our approach. Another direction is to integrate fault-tolerance techniques in order to automatically resolve abnormal proxy-disconnection situations. Other issues that will need addressing concern the definition of a performance model and analysis tool to take various network parameters into consideration.

REFERENCES

Abonamah, A., Al-Rawi, A., & Minhaz, M. (2003). A unifying Web caching architecture for the WWW.Zayed University, Abu Dhabi, UAE. *IEEE ISSPIT, The IEEE Symposium on Signal Processing and Information Technology*, Maroc, January (pp. 82-94).

Acharya, S., & Smith, B. (2000). MiddleMan: A video caching proxy server. In *Proceedings of the IEEE NOEEDAV, The 10th International Workshop on Network and Operating System Support for Digital Audio and Video*, Chapel Hill, North Carolina, USA.

Armon, A., & Levy, H. (2003). Cache satellite distribution systems: Modeling and analysis. In *IEEE INFOCOM, Conference on Computer Communications*, San Francisco, California.

Chankhunthod, A., Danzig, P., Neerdaels, P., Schwartz, M., & Worrell, K. (1996). Hierarchical Internets object cache. *Proceedings of the USENIX Technical Conference.*

Chu, Y. H., Rao, S., & Zhang, H. (2000). A case for end system multicast. *Proceedings of ACM Sigmetrics*, 1-12.

Fan, L., Cao, P., Almeida, J., & Broder, A. Z. (2000). Summary cache: A scalable wide-area Web cache sharing protocol. *IEEE/ACM Transactions on Networking, 8*(3), 281-293.

Francis, P., Jamin, S., Jin, C., Jin, Y., Raz, D., Shavitt, Y., & Zhang, L. (2001). Global Internet host distance estimation service. *IEEE/ACM Transactions on Networking, 9*(5), 525-540.

Guo, K., Buddhikot, M. M., & Chae, Y. (2001). RCache: Design and analysis of scalable, fault tolerant multimedia stream caching schemes. In *Proceedings of SPIE, Conference on Scalability and Traffic Control in IP Network*, Boston, Massachusetts, August.

Hofmann, M., Eugene Ng, T. S., Guo, K., Paul, S., & Zhang, H. (1999). *Caching techniques for streaming multimedia over the Internet* (Tech. Rep. No. BL011345-990409-04TM). Bell Laboratories.

Jannotti, J., Gifford, D. K., Johnson, K.L., Kaashoek, M.F., & O'Toole, J.M. (2001). Overcast: Reliable multicasting with an overlay network. *Proceedings of the Fourth Symposium on Operating System Design and Implementation (OSDI)*, 197-212.

Kangasharju, J., Hartanto, F., Reisslein, M., & Ross Keith, W. (2001). Distributing layered encoded video through caches. *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, 1791-1800.

Paul, S., & Fei, Z. (2001). Distributed caching with centralized control. *Computer Communications Journal, 24*(2), 256-268.

Povey, D., & Harrison, J. (1997). A distributed Internet cache. *Proceedings of the 20th Australian Computer Science Conference*, 175-184.

Rejaie, R., Handley, M., Yu, H., & Estrin, D. (1999). Proxy caching mechanism for multimedia playback stream in the Internet. *The Fourth International Web Caching Workshop*, San Diego, California.

Rejaie, R., & Kangasharju, J. (2001). Mocha: A quality adaptive multimedia proxy cache for Internet streaming. *Proceedings of the International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'01)*, 3-10.

Squid Web Proxy Cache. (2004). Retrieved from *http://www.squid-cache.org/*

Wang, B., Sen, S., Adler, M., & Towsley, D. (2002). Optimal proxy cache allocation for efficient streaming media distribution. *Proceedings of the Conference on Computer Communications (IEEE Infocom), 3*, 1726-1735.

Wu, K., Yu, P. S., & Wolf, J. L. (2001). Segment-based proxy caching of multimedia streams. *Proceedings of the 10th International World Wide Web Conference*, 36-44.

KEY TERMS

Cache: A disk space used to store the documents loaded from the server for future use.

Clip: A set of segments having the same salient objects.

Global Popularity: Depends on the number of requests to the object.

Local Popularity: Depends on the number of requests to a segment.

Popularity: Indicates the importance of a multimedia object. The most requested objects are the most popular. As a multimedia object is a set of segments, we distinguish two types of popularity: global popularity and local popularity.

Profile: Describes the user's interests and preferences in terms of themes (sports, cinema, news, etc.).

Proxy: A server that sits between the client and the real server. It intercepts all queries sent to the real server to see if it can fulfill them itself. If not, it forwards the query to the real server.

Proxy Cache: A proxy having a cache.

Routing Table by Profile (RTP): It contains a list of couples (treatment proxy, profile) that allows a proxy cache to choose, in function of a profile, the treatment proxy to which queries will be sent.

Start-Up Latency: The time that the user waits to start the clip display.

Treatment Proxy: Identified as a powerful machine with high capacities for storage, treatment, and/or communication. In addition to storage capacity, its main task is to manage a set of cache proxies of the same profile and to maintain the index of their content.

Unicast Stream: A data flow communicated over a network between a single sender and a single receiver.

Multimedia Technologies in Education

Armando Cirrincione
SDA Bocconi School of Management, Italy

WHAT ARE MULTIMEDIA TECHNOLOGIES

MultiMedia Technologies (MMT) are all that kind of technological tools that make us able to transmit information in a very large meaning, transforming information into knowledge through stimulating the cognitive schemes of learners and leveraging the learning power of human senses. This transformation can acquire several different forms: from digitalized images to virtual reconstructions, from simple text to iper-texts that allow customized, fast, and cheap research within texts; from communications framework like the Web to tools that enhance all our sense, allowing complete educational experiences (Piacente, 2002b).

MMT are composed by two great conceptually different frameworks (Piacente, 2002a):

- **Technological supports, as hardware and software:** all kinds of technological tools such as mother boards, displays, videos, audio tools, databases, communications software and hardware, and so on;
- **Contents:** information and to knowledge transmitted with MMT tools. Information are simply data (such as visiting timetable of museum, cost of tickets, the name of the author of a picture), while knowledge comes from information *elaborated in order to get a goal*. For instance, a complex ipertext about a work of art, where much information is connected in a logical discourse, is knowledge. For the same reason, a virtual reconstruction comes from knowledge about the rebuilt facts.

It's relevant to underline that to some extent technological supports represent a condition and a limit for contents (Wallace, 1995). In other words, content could be expressed just through technological supports, and this means that content has to be made in order to fit for specific technological support and that the limits of a specific technological support are also the limits of its content. For instance the specific architecture of a database represents a limit within which contents have to be recorded and have to be traced. This is also evident thinking about content as a communicative action: communication is strictly conditioned by the tool we are using.

Essentially, we can distinguish between two areas of application of MMT (Spencer, 2002) in education:

1. Inside the educational institution (schools, museums, libraries), with regard to all tools that foster the value of lessons or visiting during time they takes place. Here we mean "enhancing" as enhancing moments of learning for students or visitors: hypertexts, simulation, virtual cases, virtual reconstructions, active touch-screen, video, and audio tools;
2. In respect of outside the educational institution, this is the case of communication technologies such as Web, software for managing communities, chats, forums, newsgroups, for long-distance sharing materials, and so on. The power of these tools lies on the possibilities to interact and to cooperate in order to effectively create knowledge, since knowledge is a social construct (Nonaka & Konno, 1998; von Foester, 1984; von Glaserfeld, 1984).

Behind these different applications of MMT lies a common database, the heart of the multimedia system (Pearce, 1995). The contents of both applications are contained into the database, and so the way applications can use information recorded into database is strictly conditioned by the architecture of database itself.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

DIFFERENT DIMENSIONS OF MMT IN TEACHING AND LEARNING

We can distinguish two broader framework for understanding contributions of MMT to teaching and learning.

The first pattern concerns the place of teaching; while in the past, learning generally required the simultaneous presence of teacher and students for interaction, now it is possible to teach long distance, thanks to MMT.

The second pattern refers to the way people learn; they can be passive or they can interact. The interaction fosters learning process and makes it possible to generate more knowledge in less time.

Teaching on Site and Distance Teaching

Talking about MMT applications in education requires to separate learning on-site and distance learning, although both are called e-learning (electronic learning). E-learning is a way of fostering learning activity using electronic tools based on multimedia technologies (Scardamaglia & Bereiter, 1993).

The first pattern generally uses MMT tools as a support to traditional classroom lessons; the use of videos, images, sounds, and so on can dramatically foster the retention of contents in student's minds (Bereiter, Scardamaglia, Cassels, & Hewitt, 1997).

The second pattern, distance teaching, requires MMT applications for a completely different environment, where students are more involved in managing their commitment. In other words, students in e-learning have to use MMT applications more independently than they are required to do during a lesson on site. Although this difference is not so clear among MMT applications in education, and it is possible to get e-learning tools built as they had to be used during on-site lessons and vice-versa, it is quite important to underline the main feature of e-learning not just as a distant learning but as a more independent and responsible learning (Collins, Brown, & Newman, 1995).

There are two types of distance e-learning: self-paced and leader-led. The first one is referred to the process students access computer-based (CBT) or Web-based (WBT) training materials at their own pace. Learners select what they wish to learn and decide when they will learn it.

The second one, leader-led e-learning, involves an instructor and learners can access real-time materials (synchronous) via videoconferencing or audio or text messaging, or they can access delayed materials (asynchronous).

Both the cited types of distance learning use performance support tools (PST) that help students in performing a task or in self-evaluating.

Passive and Interactive Learning

The topic of MMT applications in an educational environment suggests distinguishing two general groups of applications referring to required students behaviour: passive or interactive. Passive tools are ones teachers use just to enhance the explanation power of their teaching: videos, sounds, pictures, graphics, and so on. In this case, students do not interact with MMT tools; that means MMT application current contents don't change according to the behaviour of students.

Interactive MMT tools change current contents according to the behaviour of students; students can chose to change contents according with their own interests and levels. Interactive MMT tools use the same pattern as the passive ones, such as videos, sounds, and texts, but they also allow the attainment of special information a single student requires, or they give answers just on demand. For instance, self-evaluation tools are interactive applications. Through interacting, students can foster the value of time they spent in learning, because they can use it more efficiently and effectively.

Interaction is one of the most powerful instruments for learning, since it makes possible active cooperation in order to build knowledge. Knowledge is always a social construct, a sense-making activity (Weick, 1995) that consists in giving meaning to experience. Common sense-making fosters knowledge building thanks to the richness of experiences and meanings people can exchange. Everyone can express his own meaning for an experience, and interacting this meaning can be elaborated and it can be changed until it becomes common knowledge. MMT help this process since they make possible interaction in less time and over long distance.

THE LEARNING PROCESS BEHIND E-LEARNING

Using MMT applications in education allows to foster learning process since there are several evidences that people learn more rapidly and deeply from words, images, animations, and sounds, than from words alone (Mayer, 1989; Mayer and Gallini, 1990). For instance, in the museum sector there is some evidence of the effectiveness of MMT devices too: Economou (1998) found firstly that people spend more time and learn more within a museum environment where there are MMT devices.

The second reason why MMT fosters learning derives from interaction they make possible. MMT allow building a common context of meaning, to socialize individual knowledge, to create a network of exchanges among teacher and learners. This kind of network is more effective when we consider situated knowledge, a kind of knowledge adults require that is quite related to problem-solving.

Children and adults have different pattern of learning, since adults are more autonomous in the learning activity and they also need to refer new knowledge to the old one they possess. E-learning technologies have developed a powerful method in order to respond more effectively and efficiently to the needs of children and adults: the "learning objects" (LO). Learning objects are single, discrete modules of educational contents with a certain goal and target. Every learning object is characterized by content and a teaching method that foster a certain learning tool: intellect, senses (sight, heard, and so on), fantasy, analogy, metaphor, and so on. In this way, every learner (or every teacher, for children) can choose its own module of knowledge and the learning methods that fit better with his own level and characteristics.

As far as the reason why people learn more with MMT tools, it is useful to consider two different theories about learning: the *information delivery theory* and the *cognitive theory*. The first one stresses teaching as just a delivery of information and it looks at students as just recipients of information.

The second one, the cognitive theory, considers learning as a sense-making activity and teaching as an attempt to foster appropriate cognitive processing in the learner. According to this theory, instructors have to enable and encourage students to actively process information: an important part of active processing is to construct pictorial and verbal representations of the lesson's topics and to mentally connect them. Furthermore, archetypical cognitive processes are based on senses, that means: humans learn immediately with all five senses, elaborating stimuli that come from environment. MTT applications can be seen as virtual reproductions of environment stimuli, and this is another reason why MMT can dramatically fostering learning through leveraging senses.

CONTRIBUTIONS AND EFFECTIVENESS OF MMT IN EDUCATION

MMT allow transferring information with no time and space constraints (Fahy, 1995).

Space constraints refer to those obstacles that arise from costs of transferring from one place to another. For instance, looking at a specific exhibition of a museum, or a school lesson, required to travel to the town where it happens; participating to a specific meeting or lesson that takes place in a museum or a school required to be there; preparing an exhibition required to meet work group daily. MMT allows the transmission of information everywhere very quickly and cheaply, and this can limit the space-constraint; people can visit an exhibition stay at home, just browsing with a computer connected on internet. Scholars can participate to meeting and seminars just connecting to the specific web site of the museum. People who are organizing exhibitions can stay in touch with the Internet, sending to each other their daily work at zero cost.

Time constraint has several dimensions: it refers to the need to catch something just when it takes place. For instance, a lesson requires to be attended when it takes place, or a temporary exhibition requires to be visited during the days it's open and just for the period it will stay in. For the same reason, participating in a seminar needs to be there when it takes place.

But time constraint refers also to the limits people suffer in acquiring knowledge: people can pay attention during a visit just for a limited period of time, and this is a constraint for their capability of learning about what they're looking for during the visiting.

Another dimension of time constraint refers to the problem of rebuilding something that happened in the

past; in the museum sector, it is the case of extemporary art (body art, environmental installations, and so on) or the case of an archaeological site, and so on.

MMT help to solve these kinds of problems (Crean, 2002; Dufresne-Tassé, 1995; Sayre, 2002) by making it possible:

- to attend school lessons on the Web, using videostreamer or cd-rom, allowing repetition of the lesson or just one difficult passage of the lesson (solving the problem of decreasing attention over time);
- to socialize the process of sense making, and so to socialized knowledge, creating networks of learners;
- to prepare the visit through a virtual visit to the Web site: this option allows knowing ex-ante what we are going to visit, and doing so, allows selection of a route more quickly and simply than a printed catalogue. In fact, thanks to ipertext technologies, people can obtain lot of information about a picture just when they want and just as they like. So MMT make it possible to organize information and knowledge about heritage into databases in order to customize the way of approaching cultural products. Recently the Minneapolis Institute of Art has started a new project on Web, projected by its Multimedia department, that allow consumers to get all kind of information to plan a deep organized visit;
- to cheaply create different routes for different kind of visitors (adults, children, researcher, academics, and so on); embodying these routes into high tech tools (PCpalm, LapTop) is cheaper than offering expensive and not so effective guided tours.
- to re-create and record on digital supports something that happened in the past and cannot be renewed. For instance the virtual re-creation of an archaeological site, or the recording of an extemporary performance (so diffuse in contemporary art).

For all the above reasons, MMT enormously reduces time and space constraints, therefore stretching and changing the way of teaching and learning.

REFERENCES

Bereiter C., Scardamalia M., Cassels C., & Hewitt J. (1997). Postmodernism, knowledge building and elementary sciences, *The Elementary School Journal*, *97*(4), 329-341.

Collins, A., Brown, J.S., & Newman S. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. In Resnick, L.B. (Ed.), *Cognition and instructions: Issues and agendas*, Lawrence Erlbaum Associates.

Crean B. (2002). Audio-visual hardware. In B. Lord & G.D. Lord (Eds.), *The manual of museum exhibitions*, Altamira Press.

Dufresne-Tassé, C. (1995). Andragogy (adult education) in the museum: A critical analysis and new formulation. In E. Hooper-Greenhill (Ed.), *Museum, media, message*, London: Routledge.

Economou, M. (1998). The evaluation of museum multimedia applications: Lessons from research. *Museum Management and Curatorship*, *17*(2), 173-187.

Fahy, A. (1995). *Information, the hidden resources, museum and the Internet.* Cambridge: Museum Documentation Association.

Mayer, R.E. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology, 81*(2), 240-246.

Mayer, R.E. & Gallini, J.K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology*, *82*(4), 715-726.

Nonaka, I. & Konno, N. (1998). The concept of Ba: Building a foundation for knowledge creation. *California Management Review*, *40*(3), 40-54.

Pearce, S. (1995). Collecting as medium and message. In E. Hooper-Greenhill (Ed.), *Museum, media, message.* London: Routledge.

Piacente, M. (2002a) Multimedia: Enhancing the experience. In B. Lord & G.D. Lord (Eds.), *The manual of museum exhibitions.* Altamira Press.

Piacente, M. (2002b). The language of multimedia. In B. Lord & G.D. Lord (Eds.), *The manual of museum exhibitions.* Altamira Press.

Sayre, S. (2002). Multimedia investment strategies at the Minneapolis Institute of Art. In B. Lord & G.D. Lord (Eds.), *The manual of museum exhibitions.* Altamira Press.

Spencer, H.A.D. (2002). Advanced media in museum exhibitions. In B. Lord & G.D. Lord (Eds.), *The manual of museum exhibitions.* Altamira Press.

Von Foester, H. (1984). Building a reality. In P. Watzlawick (Ed.), *Invented reality*. New York: WWNorton & C.

Von Glaserfeld, E. (1984). Radical constructivism: An introduction. In P. Watzlawick (Ed.) *Invented Reality*. New York: WWNorton & C.

Wallace, M. (1995). Changing media, changing message. In E. Hooper-Greenhill (Ed.), *Museum, media, message.* London: Routledge.

Watzlawick, P. (Ed.) (1984). *Invented reality*. New York: WWNorton & C.

Weick, K. (1995). *Sensemaking in organizations.* Thousand Oaks, CA: Sage Publications.

KEY TERMS

CBT: Computer based training; training material is delivered using hard support (CDRom, films, and so on) or on site.

Cognitive Theory: Learning as a sense-making activity and teaching as an attempt to foster appropriate cognitive processing in the learner.

E-Learning: A way of fostering learning activity using electronic tools based on multimedia tecnologies.

Information Delivery Theory: Teaching is just a delivery of information and students are just recipients of information.

Leader-Led E-Learning: Electronic learning that involves an instructor and where students can access real-time materials (synchronous) via videoconferencing or audio or text messaging, or they can access delayed materials (asynchronous).

LO: Learning objects; single, discrete modules of educational contents with a certain goal and target, characterized by content and a teaching method that foster a certain learning tool: intellect, senses (sight, heard, and so on), fantasy, analogy, metaphor, and so on.

MMT: Multimedia technologies; all technological tools that make us able to transmit information in a very large meaning, leveraging the learning power of human senses and transforming information into knowledge stimulating the cognitive schemes of learners.

PST: Performance support tools; software that helps students in performing a task or in self-evaluating.

Self Paced E-Learning: Students access computer based (CBT) or Web-based (WBT) training materials at their own pace and so select what they wish to learn and decide when they will learn it.

Space Constraints: All kind of obstacles that arise costs of transferring from a place to another.

Time Constraints: It refers to the need to catch something just when it takes place because time flows.

WBT: Web-based training; training material is delivered using the World Wide Web.

The N-Dimensional Geometry and Kinaesthetic Space of the Internet

Peter Murphy
Victoria University of Wellington, New Zealand

INTRODUCTION

What does the space created by the Internet look like? One answer to this question is to say that, because this space exists "virtually", it cannot be represented. The idea of things that cannot be visually represented has a long history, ranging from the romantic sublime to the Jewish God. A second, more prosaic, answer to the question of what cyberspace looks like is to imagine it as a diagram-like web. This is how it is represented in "maps" of the Internet. It appears as a mix of cross-hatching, lattice-like web figures, and hub-and-spoke patterns of intersecting lines.

This latter representation, though, tells us little more than that the Internet is a computer-mediated network of data traffic, and that this traffic is concentrated in a handful of global cities and metropolitan centres. A third answer to our question is to say that Internet space looks like its representations in graphical user interfaces (GUIs). Yet GUIs, like all graphical designs, are conventions. Such conventions leave us with the puzzle: are they adequate representations of the nature of the Net and its deep structures?

Let us suppose that Internet space can be visually represented, but that diagrams of network traffic are too naïve in nature to illustrate much more than patterns of data flow, and that GUI conventions may make misleading assumptions about Internet space, the question remains: what does the structure of this space actually look like? This question asks us to consider the intrinsic nature, and not just the representation, of the spatial qualities of the Internet. One powerful way of conceptualising this nature is via the concept of hyperspace.

The term hyperspace came into use about a hundred years before the Internet (Greene, 1999; Kaku, 1995; Kline, 1953; Rucker, 1984; Rucker, 1977; Stewart, 1995; Wertheim, 1999). In the course of the following century, a number of powerful visual schemas were developed, in both science and art, to depict it. These schemas were developed to represent the nature of four-dimensional geometry and tactile-kinetic motion—both central to the distinctive time-space of twentieth-century physics and art. When we speak of the Internet as hyperspace, this is not just a flip appropriation of an established scientific or artistic term. The qualities of higher-dimensional geometry and tactile-kinetic space that were crucial to key advances in modern art and science are replicated in the nature and structure of space that is browsed or navigated by Internet users. Notions of higher-dimensional geometry and tactile-kinetic space provide a tacit, but nonetheless powerful, way of conceptualising the multimedia and search technologies that grew up in connection with networked computing in the 1970s to 1990s.

BACKGROUND

The most common form of motion in computer-mediated space is via links between two-dimensional representations of "pages". Ted Nelson, a Chicago-born New Yorker, introduced to the computer world the idea of linking pages (Nelson, 1992). In 1965 he envisaged a global library of information based on hypertext connections. Creating navigable information structures by hyper-linking documents was a way of storing contemporary work for future generations. Nelson's concept owed something to Vannevar Bush's 1945 idea of creating information trails linking microfilm documents (Bush, 1945). The makers of HyperCard and various CD-Rom stand-alone computer multimedia experiments took up the hypertext idea in the 1980s. Nelson's concept realized its full potential with Berners-Lee's design for the "World Wide Web" (Berners-Lee, 1999). Berners-Lee worked out the simple, non-proprietary protocols required to effectively fuse hyper-linking with self-organized computer networking. The result was hyper-linking

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

between documents stored on any Web server anywhere in the world.

The hyper-linking of information-objects (documents, images, sound files, etc.) permitted kinetic-tactile movement in a virtual space. This is a space—an information space—that we can "walk through" or navigate around, using the motor and tactile devices of keyboards and cursors, and motion-sensitive design cues (buttons, arrows, links, frames, and navigation bars). It includes two-dimensional and three-dimensional images that we can move and manipulate. This space has many of the same characteristics that late nineteenth century post-Euclidean mathematicians had identified algebraically, and that early 20th-century architects and painters set out to represent visually.

The term hyperspace came into use at the end of the 19th century to describe a new kind of geometry. This geometry took leave of a number of assumptions of classical or Euclidean geometry. Euclid's geometry assumed space with flat surfaces. Nicholas Lobatchevsky and Bernhard Riemann invented a geometry for curved space. In that space Euclid's axiom on parallels no longer applied. In 1908, Hermann Minkowski observed that a planet's position in space was determined not only by its x, y, z coordinates but also by the time it occupied that position. The planetary body moved through space in time. Einstein later wedded Minkowski's hyperspace notion of space-time to the idea that the geometry of planetary space was curved (Greene, 1999; Hollingdale, 1991; Kline, 1953).

Discussion of hyperspace and related geometric ideas signalled a return to the visualization of geometry (Kline, 1953). Ancient Greeks thought of geometry in visual terms. This was commonplace until Descartes' development of algebra-based geometry in the 17th century. Euclidean geometry depicted solids in their three dimensions of height, width, and breadth. The 17th century coordinate geometry of René Descartes and Pierre Fermat rendered the visual intuitions of Euclid's classical geometry into equations—that is, they translated the height, depth, and breadth of the x, y, z axes of a three-dimensional object into algebra. In contrast, in the 20th century, it was often found that the best way of explaining post-Euclidean geometry was to visually illustrate it.

This "will to illustrate" was a reminder of the traditionally close relationship between science and art. Mathematics was common to both. It is not surprising then that post-Euclidean geometry was central not only to the new physics of Einstein and Minkowski but also to the modern art of Cézanne, Braque, and Picasso (Henderson, 1983). In turn, the visualised geometry of this new art and science laid the basis for the spatial intuitions that regulate movement and perception in Internet-connected multimedia environments. In geometric terms, such environments are "four dimensional". In aesthetic terms, such environments have a "cubist" type of architecture.

Technologies that made possible the navigable medium of the Internet—such as the mouse, the cursor, and the hypertext link—all intuitively suppose the spatial concepts and higher dimensional geometries that typify Cézanne-Picasso's multi-perspective space and Einstein-Minkowski's space-time. The central innovation in these closely related concepts of space was the notion that space was not merely visual, but that the visual qualities of space were also tactile and kinetic. Space that is tactile and kinetic is fundamentally connected to motion, and motion occurs in time. Space and time are united in a continuum. The most fundamental fact about Internet or virtual space is that it is not simply space for viewing. It is not just "space observed through a window". It is also space that is continually touched—thanks to the technology of the mouse and cursor. It is also space that is continually moved through—as users "point-and-click" from link to link, and from page to page. Consistent with the origins of the term, the hyperspace of the Internet is a form of space-time: a type of space defined and shaped by movement in time—specifically by the motions of touching and clicking.

CRITICAL ISSUES

When we look at the world, we do so in various ways. We can stand still, and look at scenes that either move across our visual field or are motionless. When we do this, we behave as though we were "looking through a window". The window is one of the most powerful ways we have for defining our visual representations. The aperture of a camera is like a window. When we take a picture, the window-like image is frozen in time. The frame of a painting functions in the same way. Whether the scene depicted obeys the laws of

perspective or not, the viewer of such paintings is defined (by the painting itself) as someone who stands still and observes. Even film—the moving picture—normally does not escape this rule. Its succession of jump cut images are also a series of framed images.

Windows and window-frame metaphors dominate GUI design. Graphical user interfaces enabled the transition from command-line to visual processing of information. From their inception, GUIs were built on the metaphor of windows. Ivan Sutherland at MIT conceived the GUI window in the early 1960s—for a computer drawing-program. Douglas Engelbart reworked the idea to enable multiple windows on a screen. Alan Kay, at Xerox's Palo Alto Center, devised the mature form of the convention—overlapping windows—in 1973 (Gelernter, 1998; Head, 1999).

"Looking through a window", however, is not the only kind of visual experience we have. Much of our looking is done "on the move". Sometimes we move around still objects. This experience can be represented in visual conventions. Many of Cézanne's paintings, for example, mimic this space-time experience (Loran, 1963). They are composed with a still object in the centre while other objects appear to circulate around that still centre. Motion is suggested by the titling the axes of objects and planes. What the artist captures is not the experience of looking through a window into the receding distance—the staple of perspective painting—but the experience of looking at objects that move around a fixed point as if the observer was on the move through the visual field.

Sometimes this navigational perspective will take on a "relativistic" character—as when we move around things as they move around us. The visual perceptions that arise when we "walk-through" or navigate the world is quite different from the frozen moment of the traditional snap-shot. In conventional photography we replicate the sensation of standing still and looking at a scene that is motionless. In contrast, imagine yourself taking a ride on a ferryboat, and you want to capture in a still photo the sense of moving around a harbour. This is very hard to do with a photographic still image.

The development of the motion camera (for the movies) at the turn of the 20th century extended the capabilities of the still camera. A statically positioned motion camera was able to capture an image of objects moving in the cinematographer's visual field. The most interesting experiments with motion pictures, however, involved a motion camera mounted on wheels and tracks. Such a camera could capture the image of the movement of the viewer through a visual field, as the viewer moved in and around two- and three-dimensional (moving and static) objects. This was most notable in the case of the tracking shot—where the camera moves through space following an actor or object.

It was the attempt to understand this kind of moving-perception (the viewer on the move) that led to the discovery of the idea of hyperspace. Those who became interested in the idea of moving-perception noted that conventional science and art assumed that we stood still to view two-dimensional planes and three-dimensional objects. But what happened when we started to move? How did movement affect perception and representation?

It was observed that movement occurs in time, and that the time "dimension" had not been adequately incorporated into our conventional images of the world. This problem—the absence of time from our representations of three-dimensional space—began to interest artists (Cézanne) and mathematicians (Poincaré and Minkowski). Out of such rethinking emerged Einstein's theories.

Artists began to find visual ways of representing navigable space. This is a kind of space that is not only filled with static two- or three-dimensional objects that an observer views through a window. It is also space in which both observers and things observed move around. This space possesses a "fourth" dimension, the "dimension" of time. In such space, two- and three-dimensional objects are perceived and represented in distinctive ("hyper-real" or "hyper-spatial") ways.

The painters Cézanne, Picasso, and Braque portrayed the sequential navigation/rotation of a cube or other object as if it was happening in the very same moment (simultaneously) in the visual space of a painting. Imagine walking around a cube, taking successive still photos of that circumnavigation, and then pasting those photos into a single painted image. Picasso's contemporary, the Amsterdam painter-architect Theo Van Doesburg, created what he called "moto-stereometrical" architecture—three-dimensional buildings designed to represent the dimension of time (or motion). Doesburg did not just design a space that could be navigated but also a representation of how our brain perceives a building (or its

geometry) as we walk round it. Doesburg's hyperspace was composed of three-dimensional objects interlaced with other three-dimensional objects. This is a higher-dimensional analogue of the traditional Euclidean idea of a two-dimensional plane being joined to another two-dimensional plane to create a three-dimensional object. A hypersolid is a three-dimensional solid bounded by other three-dimensional solids. This type of architecture captures in one image (or one frozen moment) the navigation of objects in time.

In 1913, the New York architect, Claude Bragdon, developed various "wire diagrams" (vector diagrams) with coloured planes to represent this interlacing of three-dimensional objects. The same idea interlacing of three-dimensional object-shapes also appears in the architecture of the great twentieth-century philosopher Ludwig Wittgenstein, in the villa that he designed for his sister in Vienna in 1926 (Murphy & Roberts, 2004). Wittgenstein's contemporary, the Russian artist Alexandr Rodchenko, envisaged space as composed of objects within objects. On the painters' two-dimensional canvas, he painted circles within circles, hexagons within hexagons. If you replace the two-dimensional circle with the three-dimensional sphere, you get a hyperspace of spheres within spheres.

Hypersolids are objects with more than three dimensions [= n dimensions]. One way of thinking about hypersolids is to imagine them as "three-dimensional objects in motion" (a car turning a corner) or "three-dimensional objects experienced by a viewer in motion" (the viewer standing on the deck of a boat in motion watching a lighthouse in the distance). The hypersolid is a way of representing what happens to dimensionality (to space and our perceptions of that space) when a cube, a cone, or any object is moved before our eyes, or if we move that object ourselves, or if we move around that object (Murphy, 2001).

Consider an object that moves—because of its own motion, or because of our motion, or both. Imagine that object captured in a sequence of time-lapse photos, which are then superimposed on each other, and then stripped back to the basics of geometric form. What results from this operation is an image of a hypersolid, and a picture of what hyperspace looks like. Hyperspace is filled with intersecting, overlapping, or nested three-dimensional solids.

In the case of the navigable space of hyperlinked pages (Web pages), the perception of hyperspace remains largely in the imagination. This is simply because (to date) graphical user interfaces built to represent Web space mostly assume that they are "windows for looking through". Internet and desktop browsing is dominated by the visual convention of looking through a "window" at two-dimensional surfaces. Browsing the Net, opening files, and reading documents all rely on the convention of window-framed "pages". The mind, fortunately, compensates for this two dimensionality. Much of our three-dimensional representation of the world, as we physically walk through it, is composed in our brain. The brain creates a third dimension out of the two-dimensional plane image data that the eyes perceive (Sacks, 1995). The same thing happens to plane images when we click through a series of pages. While the pages are two-dimensional entities defined by their width and height, through the haptic experience of pointing and clicking and the motion of activating links, each two-dimensional page/plane recedes into an imaginary third dimension (of depth). Moving from one two-dimensional plane to another stimulates the imagination's representation of a third dimension. Our brain illusionistically creates a perception of depth—thus giving information an object-like 3D character. But linking does more than this. It also allows movement around and through such information objects, producing the implied interlacing, inter-relating, and nesting of these virtual volumes.

Hyperspace is a special kind of visual space. It is governed not only by what the viewer sees but also by the tactile and motor capacity of the viewer and the motion of the object observed. The tactile capacity of observers is their capacity for feeling and touching. The motor capacity of the viewer is their power to move limbs, hands, and fingers. Tactile and motor capacities are crucial as a person moves through space or activates the motion of an object in space. So it is not surprising that we refer to the "look and feel" of web sites. This is not just a metaphor. It refers to the crucial role that the sense of "feel"—the touch of the hand on the mouse—plays in navigating hyperspaces.

In hyperspace, the viewers' sight is conditioned by the viewers' moving of objects in the visual field (for example, by initiating roll-overs, checking boxes, dropping down menus, causing icons to blink), or

alternatively by the viewer moving around or past objects (for example, by scrolling, gliding a cursor, or clicking). Yet, despite such ingenious haptic-kinetic structures, the principal metaphor of GUI design is "the window". The design of navigable web space persistently relies on the intuitions of pre-Riemann space.

Consequently, contemporary GUI visual conventions only play a limited role in supplementing the mind's representation of the depth, interlacing, and simultaneity of objects. Whatever they "imagine", computer users "see" a flat world. GUI design for instance gives us an unsatisfying facsimile of the experience of "flicking through the leaves of a book". The depth of the book-object, felt by the hand, is poorly simulated in human-computer interactions. The cursor is more a finger than a hand. Reader experience correspondingly is impoverished. Beyond hyper textual links, there are to date few effective ways of picturing the interlacing of tools and objects in virtual space. The dominant windows metaphor offers limited scope to represent the simultaneous use of multiple software tools—even though 80 percent of computer users employ more than one application when creating a document.

Similar constraints apply to the representation of relations between primary data, metadata, and procedural data—or between different documents, files, and Web pages open at the same time. Overlapping windows have a limited efficacy in these situations. Even more difficult is the case where users want to represent multiple objects that have been created over time for example as part of a common project or enterprise. The metaphor of the file may allow users to collocate these objects. But we open a file just like we open a window—by looking into the flatland of 2D page-space.

CONCLUSION

While the brain plays a key role in our apprehension of kinetic-tactile n-dimensional space, the creation of visual representations or visual conventions to represent the nature of this space remains crucial. Such representations allow us to reason about, and explore, our intuitions of space-time. In the case of Internet technologies, however, designers have largely stuck with the popular but unadventurous "windows" metaphor of visual perception. The advantage of this is user comfort and acceptance. "Looking through a window" is one of the easiest to understand representations of space, not least because it is so pervasive. However, the windows metaphor is poor at representing movement in time and simultaneity in space. All of this suggests that GUI design is still in its infancy. The most challenging twentieth-century art and science gives us a tempting glimpse of where interface design might one day venture.

REFERENCES

Berners-Lee, T. (1999). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. New York: HarperCollins.

Bush, V. (1945). As we may think. *The Atlantic Monthly*, No. 176.

Floridi, L. (1999). *Philosophy and computing*. London: Routledge.

Gelernter, D. (1998). *The aesthetics of computing.* London: Weidenfeld & Nicolson.

Greene, B. (1999). *The elegant universe.* New York: Vintage.

Head, A. (1999). *Design wise: A guide for evaluating the interface design of information resources.* Medford, NJ: Information Today.

Henderson, L. (1983). *The fourth dimension and non-Euclidean geometry in modern art.* Princeton, NJ: Princeton University Press.

Hollingdale, S. (1991 [1989]). *Makers of mathematics.* Harmondsworth: Penguin.

Kaku, M. (1995 [1994]). *Hyperspace*. New York: Doubleday.

Kline, M. (1953). *Mathematics in western culture.* New York: Oxford University Press.

Loran, E. (1963 [1943]). *Cénanne's composition.* Berkeley: University of California Press.

Murphy, P. (2001). Marine Reason. *Thesis Eleven, 67*, 11-38.

Murphy P. & Roberts, D. (2004). *Dialectic of romanticism: A critique of modernism.* London: Continuum.

Nelson, T. (1992 [1981]). *Literary machines 93.1.* Watertown, MA: Eastgate Systems.

Rucker, R. (1984). *The fourth dimension.* Boston: Houghton Mifflin.

Rucker, R. (1977). *Geometry, relativity and the fourth dimension.* New York: Dover.

Sacks, O. (1995). *An anthropologist on Mars.* London: Picador.

Stewart, I. (1995 [1981]). *Concepts of modern mathematics.* New York: Dover.

Wertheim, M. (1999). *The pearly gates of cyberspace: A history of space from Dante to the Internet*. Sydney: Doubleday.

KEY TERMS

Design: The structured composition of an object, process, or activity.

Haptic: Relating to the sense of touch.

Hyperspace: Space with more than three dimensions.

Metaphor: The representation, depiction or description of one thing in terms of another thing.

Multiperspectival Space: A spatial field viewed simultaneously from different vantage points.

Virtual Space: Space that is literally in a computer's memory but that is designed to resemble or mimic some more familiar conception of space (such as a physical file or a window or a street).

Web Server: A network computer that delivers Web pages to other computers running a client browser program.

Network Intrusion Tracking for DoS Attacks

Mahbubur R. Syed
Minnesota State University, USA

Mohammad M. Nur
Minnesota State University, USA

Robert J. Bignall
Monash University, Australia

INTRODUCTION

In recent years the Internet has become the most popular and useful medium for information interchange due to its wide availability, flexibility, universal standards, and distributed architecture. As an outcome of increased dependency on the Internet and networked systems, intrusions have become a major threat to Internet users. Network intrusions may be categorized into the following major types:

- Stealing valuable and sensitive information
- Destroying or altering information
- Obstructing the availability of information by destroying the service-providing ability of a victim's server

The first two types of intrusions can generally be countered using currently available information- and security-management technologies. However, the third category has a lot more difficult and unsolved issues, and is very hard to prevent. Two very common and well-known attack approaches in this category are the following:

- **Denial-of-Service (DoS) Attacks:** In DoS attacks, legitimate users are deprived of accessing information on the targeted server since its available resources (e.g., memory, processing power) as well as network bandwidth are entirely consumed by a large number of incoming packets from attackers. The attackers can hide their true identity by forging the source IP (Internet protocol) address of the attack packets since they do not need to receive any response back from the victim.
- **Worms:** Worms are self-propagating (do not require user interaction or assistance), malicious codes. They can develop DoS attacks or change sensitive configurations.

Challenges in Network-Intrusion Tracking for DoS Attacks

According to a Computer Security Institute (CSI; 2003) and FBI survey, the total financial loss in the U.S.A. during the first quarter of 2003 due to computer-related crime, which included unauthorized insider access, viruses, insider Net abuse, telecom fraud, DoS attacks, theft of proprietary information, financial fraud, sabotage, system penetration, telecom eavesdropping, and active wiretapping, amounted to $201,797,340. The losses caused by DoS attacks were the highest, amounting to 35% of the total, and were already significantly higher than in previous years. A comparative year-by-year breakdown is shown in Table 1 (Computer Security Institute).

DoS attacks are easy to implement and yet are difficult to prevent and trace. A large amount of money and effort are spent to secure organizations from Internet intrusions.

SOME BASIC FORMS OF DOS ATTACKS

Denial-of-service attacks come in a variety of forms and target a variety of services. Attackers are continuously discovering new forms of attacks using security holes in systems and protocols. Some former

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. CSI/FBI Computer crime and security survey report (in U. S. Dollars)

Year	*2000*	*2001*	*2002*	*2003 (part)*
Total Loss	265,337,990	377,828,700	455,848,000	201,797,340
Loss due to DoS	8,247,500	4,283,600	18,370,500	65,643,300

and very basic forms of DoS attacks, such as the TCP (transmission-control protocol) SYN flood, Smurf attack, and UDP (user datagram packets) flood, are briefly outlined below to clarify the underlying concept.

In TCP SYN flooding, an adversary requests TCP connections by sending TCP SYN (TCP SYNchronization request) packets containing incorrect or nonexistent IP source addresses to the targeted victim. The victim responds with a SYN-ACK (SYNchronization ACKnowledgement) packet to the forged source IP address, but never gets a reply, which leaves the last part of a three-way handshake incomplete. Consequently, half-open connections quickly fill up the connection queue of the targeted server and it becomes unable to provide services to legitimate TCP users.

In a Smurf attack (also known as a Ping attack), the adversary broadcasts ping messages with the targeted victim's source address and multicast destination addresses to various networks. All computers in those networks consequently reply to the source address, flooding the targeted victim with pong messages that it did not request. ICMP (Internet control message protocol) flood attacks use a similar method.

In a UDP-flood attack, a large number of UDP packets are sent to the target, overwhelming available bandwidth and system resources.

Distributed Denial-of-Service Attacks

Distributed DoS (DDoS) attacks are a more powerful and more destructive variation of DoS attacks. In DDoS attacks, a multitude of compromised systems attack a single target simultaneously and hence are more malicious and harder to prevent and trace compared to DoS. The victim of DDoS attacks is not limited to the primary target; in reality all of the systems controlled and used by the intruder are victimized as well.

DEFENDING AGAINST NETWORK INTRUSION

Defense against network intrusion includes three steps: prevention, detection, and attack-source identification.

Intrusion prevention includes the following:

- **Access Control:** Firewalls control access based on the source IP address, destination IP address, protocol type, source port number, and destination port number, or based on the customer need. However, if an attacker attempts to exploit, for example, the WWW (World Wide Web) server using HTTP (hypertext transfer protocol), the firewall cannot prevent it.
- **Preventing Transmission of an Invalid Source IP Address:** Egress filtering of outgoing packets before sending them out to the Internet (i.e., discarding packets with forged IP address on the routers that connect to the Internet) would cease intrusion by outsiders immediately.
- **Increased Fault Tolerance:** Servers or any other possible victims should be well equipped to deal with network intrusions and should work even in the presence of an intrusion or when partially compromised, for example, systems with a larger connection queue to deal with TCP-SYN attacks.

- Intrusion-detection systems (IDSs) continuously monitor incoming traffic for attack signatures (features from previously known attacks). Ingress filtering is performed on the router by the IDS.
- Intrusion tracing identifies the origin of the attack using techniques such as IP traceback.

However, these defense systems are often not enough for dealing with DoS attacks. The Internet protocol does not have any built-in mechanism to ensure that the source address of an IP packet actually represents the origin of the packet. Due to the lack of knowledge of the attacker's identity, taking immediate action to stop the attack becomes impossible. Moreover, there is no built-in quality-of-service (QoS) or resource-restriction mechanism in use that can prohibit an attacker from consuming all available bandwidth. Since IP alone does not address this security issue, we need some IP-traceback technology that can identify an attack host by tracking the attack packets back to their source along the route they traveled. While current commercially available technologies are not capable of preventing DoS attacks, the ability to trace such attacks to their source can act as a deterrent. A significant amount of research is being done for more cost-effective and efficient IP-traceback techniques.

DESIGN CHALLENGES FOR EFFECTIVE IP-TRACEBACK TECHNIQUES

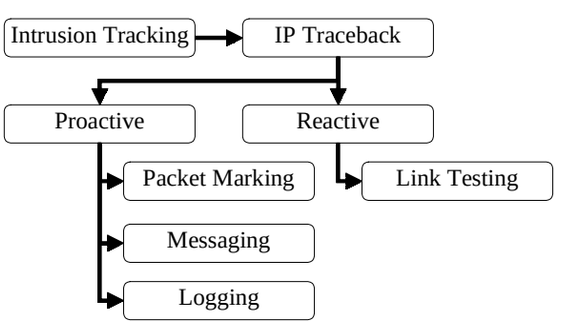
IP traceback requires an exchange of information between the routers along an attack path, so the implementation of supporting protocols throughout the Internet is critical for defending against DoS attacks (and most other network intrusions). A number of factors must be considered when designing an effective traceback mechanism.

- Attack packets may be of any type, volume, and source address.
- The attack duration may be very short or long.
- Packets may be designed to be lost, incorrectly ordered, or inserted by the attackers anywhere across the attack path in order to misguide the tracing process.
- An attack may be coming from multiple sources, for example, in DDoS attacks. These sources may simply be slaves (compromised hosts) or secondary victims of the attack.
- Existing routers are designed only for forwarding packets and have CPU (central processing unit) and memory limitations. On-the-fly processing for tracking or analyzing packets, and for information exchange need to be designed so as to avoid placing too much burden on them.
- Extracting and analyzing information from a traffic flow may generate a large amount of data that is difficult to store, maintain, and search.
- The flow of legitimate traffic must not be interrupted or delayed.
- There might be firewalls or gateways across the attack path, which may mislead or block the tracing process.
- Traceback techniques should meet the following criteria: They need to be fast, have a low cost and deployment time, minimize manual and individual configurations, and utilize existing technologies. They should also support an incremental implementation.
- Many ISPs (Internet service providers) may not be able to afford to participate or even to cooperate because of the additional costs involved, so requiring a minimal amount of assistance from other network ISPs or operators is a major consideration.

INTRUSION TRACING USING IP TRACEBACK

Designing a traceback system is extremely difficult and challenging. Figure 1 shows a classification based on a survey of different intrusion-tracing techniques from the literature.

Figure 1. Basic IP-traceback strategies



Proactive Tracing

Proactive IP-traceback techniques prepare tracking information while packets are in transit, and this is done regardless of the occurrence of an attack. If an attack takes place, the victim can use this captured information to identify the attack source. If no attacks occur, all of the time and effort put into generating the tracing information is wasted. Existing proactive tracing methods essentially follow three strategies: packet marking, messaging, and logging.

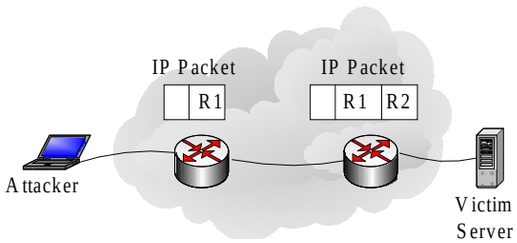
Packet Marking

In packet marking, different strategies are used to store information about the routers along the path of the traversing packets. In the event of an attack, these marked packets may be used to reconstruct the packet's travel path to its source to locate the attacker. Packet marking requires that all the routers throughout the Internet are able to mark packets. To reduce the processing time and per-packet space requirements, most such techniques use probability-based marking strategies. However, this means that a large volume of attack packets is required in order to collect sufficient information to identify the attack path. Moreover, probability-based strategies are less robust against DDoS attacks and so design improvements are needed to deal with the increasing number of DDoS attacks.

As a continuous effort of researchers to develop better and more efficient techniques, a series of proactive IP-traceback strategies such as node appending, node sampling, edge sampling, compressed edge-fragment sampling, SNITCH (simple, novel IP traceback using compressed headers), and Pi have been proposed. Each of these is, in fact, an incremental improvement on the previous one.

Node appending (Savage, Wetherall, Karlin, & Anderson, 2000) is based on the simple concept that each router appends its address to the end of the packets traversing through it. Thus, a complete and ordered list of IP addresses of the entire travel path is contained in each packet. A victim of an attack can construct the attack path very easily and quickly by examining just a single packet. Unfortunately, the per-packet space requirement is too high to be accommodated in IP-based packets, and adding IP addresses to each packet on the fly imposes a high router-processing overhead.

Figure 2: Node appending: Routers add their IP address (e.g., R1, R2) to the packets travelling through them



As an effort to overcome the high router overhead and huge per-packet space-requirement problem of node appending, the node-sampling method (Savage et al., 2000) adds only one router's address to a packet instead of storing the IP addresses of all routers along the entire travel path. Each router writes its address with some probability *p*; so, a router may overwrite some addresses written by previous routers, which means that the frequency of received marked packets from a given router decreases as the distance between that router and the victim increases. Reconstructing the attack path has become a much slower and more uncertain process due to the complexity of computing the order of the routers from the samples.

The edge-sampling method (Savage et al., 2000) aims to reduce the complexity and processing time of node sampling. It marks the edges and their distance from the victim, with a probability *p*, along the attack path. This requires 72 bits (two 32-bit IP addresses and one 8-bit distance field) of additional space in the packet header. The advantage of the distance field is that it prevents fake edge insertion by an attacker in a single-source attack because the packets sent by an attacker must have a distance greater than or equal to the length of the true attack path. However, this method incurs a high router overhead and to some extent reintroduces per-packet space-requirement problems.

Compressed edge-fragment sampling (Savage et al., 2000) stores a random fragment of the subsequent edges constructed by performing Exclusive-

OR (XOR) two adjacent nodes (e.g., a ⊕ b) in the IP identification field. The victim reconstructs the original path by XORing the received values (e.g., b ⊕ (a ⊕ b)). This method improves on the edge-sampling method by reducing the per-packet storage requirement to 16 bits. However, using the IP identification field for storing the edge causes serious conflicts with IP datagram fragmentation and some IPsec (IP security) protocols. Also, the computational complexity is greatly increased. Attackers can insert fake edges since the victim cannot distinguish between genuinely marked packets and attack packets unmarked by intermediate routers.

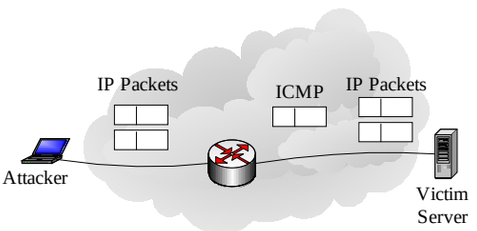
Hassan, Marcel, and Alexander (2003) propose the SNITCH protocol, which is, in fact, an effort to improve Savage et al.'s (2000) compressed edge-sampling protocol in terms of better space accommodation and more accurate attack path identification. The main feature of this probabilistic packet-marking technique is the use of header compression (similar to RFC (request for comments) 2507) in order to accommodate space for insertion of traceback information. XOR and bit rotation have been used to improve the efficiency of the method.

Pi is a path identification method for DDoS attacks proposed by Yaar, Perrig, and Song (2003). It is based on an *n*-bit scheme where a router marks an edge formed by concatenating the last *n* bits from the hash of its IP address with the previous router's IP address. The edges are stored in the IP identification field of the packets it forwards. This method also offers improvement in avoiding the overwriting of markings by routers close to the victim. The victim drops incoming packets matching the markings of identified attack packets. In some cases, however, this method cannot ensure the same and correct ID for the same path. Ahn, Wee, and Hong (2004) suggest the use of XOR after the first router overwrites its marking value to overcome this kind of shortcomings.

Messaging

A router in a messaging strategy creates messages containing information about the traversing IP packets and itself, and sends these messages to the packet's destination. The victim can construct the attack path from the messages it receives. Messaging is very similar to packet marking except that the tracking information is sent out of band, in separate packets, giving an easy and effective solution to the per-packet space-requirement problem of packet-marking protocols.

Figure 3: ICMP messaging: Routers send ICMP messages about the packets with a probability p



In ICMP-messaging method (Bellovin, 2000), ICMP traceback packets are forwarded by the router with a probability of 1:20,000 (to avoid an increase in network traffic). This should be effective for typical DoS attacks that contain thousands of attack packets per second.

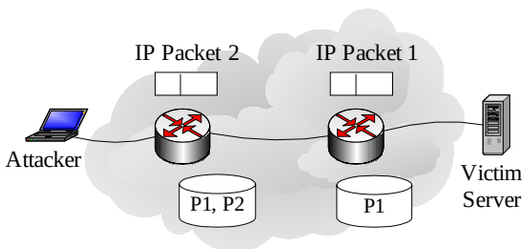
The main benefit of ICMP Messaging is its compatibility with existing protocols. However, ICMP messages are increasingly being differentiated from normal traffic due to their abuse in different attacks (Bellovin, 2000). They also increase network traffic to some extent. Distant routers contribute fewer messages compared to the closest routers in the case of DDoS attacks. Also, false and misleading ICMP traceback messages can be sent by attackers. However, with the use of encryption and a key-distribution scheme, this approach could become secure and effective (Hassan et al., 2003).

Logging

In a logging strategy, packets are logged at the routers they travel through, and then data-mining techniques are applied on that logged information to determine the path that the attack packets have traversed.

Logging is a useful strategy because it can trace an attack long after it has ended. It can handle single-packet attacks and DDoS attacks. However, this method imposes high implementation and maintenance costs mainly for an extremely large and fast

Figure 4: Routers store packet information in a database



storage capacity. To construct an attack path, the logged data must be shared among ISPs, which raise concerns about data security and privacy.

Baba and Matsuda (2003) have proposed a proactive and multicomponent distributed logging technique, where forwarding nodes store the data link layer identifier of the previous node in addition to the information about traversing IP packets. This method also proposes an overlay network containing sensors for monitoring attacks, tracers for logging malicious traffic, and monitoring managers for controlling sensors and tracers and managing the entire tracing process.

Reactive Tracing

In a reactive approach, tracing is performed after an attack is detected and, hence, no prior processing of tracking information is required. This economizes the traceback mechanism by avoiding all the preparatory processing work undertaken by proactive approaches. The disadvantage of this approach is that if the attack ceases while the reactive tracing is being undertaken, the tracing process may fail to identify the origin of the attack due to a lack of necessary tracking information. This is the main challenge for developing effective reactive traceback techniques.

Link Testing

Link testing is a mechanism for testing network links between routers in order to determine the source of the attack. If an attack is detected across a link, the tracker program logs into the upstream router for that link. This procedure is repeated recursively on the upstream routers until the attack source is reached. Most of the reactive traceback approaches rely on link testing for tracking the attack source.

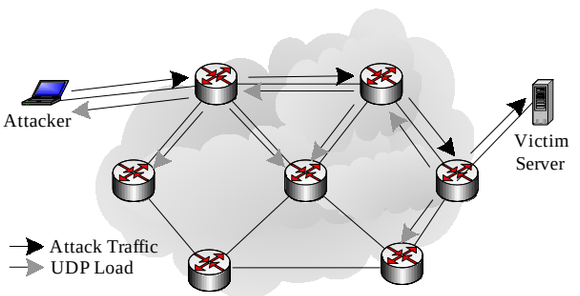
In the network ingress-filtering method proposed by Ferguson and Senie (2000), a router compares an incoming packet's source IP address with a router's routing table and discards packets with inconsistent source addresses as having been forged. This method is effective for many spoofed DoS attacks, but it can fail if an attacker changes its source IP address to one that belongs to the same network as the attacker's host.

In regular hop-by-hop tracing, the processing rapidly increases with the increase in the number of hops, and as a result, necessary tracing information might be lost or the attack may cease before the tracing process is complete. Hop-by-hop tracing with an overlay network, proposed by Stone (2000), is an effort to decrease the number of hops required for tracing with an overlay network by establishing IP tunnels between edge routers and special tracking routers, and then rerouting IP packets to the tracking routers via the tunnels.

IPsec authentication, proposed by Chang et al. (1999), is a very similar traceback technique based on the existing IPsec protocol. When an IDS detects an attack, the Internet-key-exchange (IKE) protocol establishes a tunnel between the victim and some routers in the administrative domain using IPsec security associations (SAs). If the attack continues and one of the established SA tunnels authenticates a subsequent attack packet, it is then obvious that the attack is coming from a network beyond the router at that tunnel end. This process is continued recursively until the attack source is reached. The main benefit of this approach is its compatibility with existing protocols and network infrastructure, and the fact that it does not impose any traffic overhead on the Internet. However, routers have to be synchronized and authenticated to each other, which imposes the need for worldwide collaboration. The monitoring and processing work involved puts an extra burden on ISPs because of the high resource requirements.

In controlled flooding, which is a pattern-matching-based technique, a brief burst of load consisting of packets is applied to each link attached to it using the UDP changer service, and then the change in the

Figure 5. Controlled-flooding: A burst of load is applied to attached links in order to detect the attack path



packet stream coming through that link is examined in order to decide whether that link is part of the attack path (Burch & Cheswick, 2000).

This technique does not require any support from other ISPs and is compatible with the existing network infrastructure. However, like any other reactive approach, it requires the attack to be continued until the tracing process is completed. Moreover, this approach is itself a denial-of-service attack: It puts an extremely high overhead on the routers along the attack path in order to achieve its goal. As a result, it is unsuitable and can be unethical for practical use.

CONCLUSION

There are a wide variety of intrusion techniques and there is no single solution. In fact, there is currently no effective commercial implementations available in the market to perform IP traceback effectively across the Internet in real time (Hassan et al., 2003). The few commercial traceback products that are available work only for single corporate networks and only against internal security threats.

A number of problems still exist in IP-traceback techniques, for example, deploying a working model globally, getting interactive support from all ISPs across the world, and tracing beyond firewalls and gateways in the middle of the route. A firewall that separates the network hosting the attacker from the Internet makes the attacker invisible to those outside of that network. A traceback system would require routers with higher resources to support the additional processing activities and might need hardware and/or software upgrades or replacement to ensure that packet transport occurs in close to real time. These extra cost and maintenance overheads may deter some ISPs from participating in traceback processes. Since traceback systems must be deployed throughout the Internet in order to achieve their objectives, there must be a common global policy for traceback. No single traceback technique can provide security against all types of DoS attacks. No matter what technique is adopted, there is no alternative to continued global cooperation to defeat the rapidly evolving methods of attack as no present solution will be able to stop all future security threats.

REFERENCES

Ahn, Y., Wee, K., & Hong, M. (2004). A path identification mechanism for effective filtering against DDoS attacks. *Proceedings of the Eighth Multi-Conference on Systemics, Cybernetics and Informatics, 3*, 325-330.

Baba, T., & Matsuda, S. (2003). Tracing network attacks to their sources. *Internet Computing, 6*(2), 20-26.

Bellovin, S. M. (Ed.). (2000). *ICMP traceback messages*. Internet Draft, expiriation date September, 2000. Available at *http://www.ietf.org/proceedings/01dec/I-D/draft-ietf-itrace-01.txt*

Burch, H., & Cheswick, B. (2000). Tracing anonymous packets to their approximate source. *USENIX: 14th Systems Administration Conference (LISA '00)*, 319-327.

Chang, H. Y., Narayan, R., Wu, S.F., Wang, X.Y., Yuill, J., Sargor, C., Gong, F. & Jou, F. (1999). DecIdUouS: Decentralized source identification for network-based intrusions. *Proceedings of the Sixth IFIP/IEEE International Symposium on Integrated Network Management*, 701-714.

Computer Security Institute. (2003). *2003 CSI/FBI computer crime and security survey*.

Degermark, M., Nordgren, B., & Pink, S. (1999). *IP header compression* (RFC 2507). Network Working Group.

Ferguson, P., & Senie, D. (2000). *Network ingress filtering: Defeating denial of service attacks which employ IP source address spoofing* (RFC 2827). Network Working Group.

Hassan, A., Marcel, S., & Alexander, P. (2003). IP traceback using header compression. *Computers & Security, 22*(2), 136-151.

Savage, S., Wetherall, D., Karlin, A., & Anderson, T. (2000). Practical network support for IP traceback. *Proceedings of the 2000 ACM SIGCOMM, 30*(4), 295-306.

Stone, R. (2000). *CenterTrack: An IP overlay network for tracking DoS floods*. Proceedings of the Ninth USENIX Security Symposium, Berkeley, CA.

Yaar, A., Perrig, A., & Song, D. (2003). *Pi: A path identification mechanism to defend against DDoS attacks*. Proceedings of the IEEE Symposium on Security and Privacy, California, USA.

KEY TERMS

Attack Signature: Patterns observed in previously known attacks that are used to distinguish malicious packets from normal traffic.

Egress Filtering: Process of checking whether outgoing packets contain valid source IP addresses before sending them out to the Internet. Packets with forged IP addresses are discarded on the router that connects to the Internet.

Firewall: A system that implements a set of security rules to enforce access control to a network from outside intrusions.

ICMP Message (Internet Control Message Protocol): A message control and error-reporting protocol that operates between a host and a gateway to the Internet.

IDS (Intrusion-Detection System): A utility that continuously monitors for malicious packets or unusual activity (usually checks for matches with attack signatures extracted from earlier attack packets).

Input Debugging: A process performed on a router to determine from which adjacent router the packets matching a particular attack signature are coming from.

Intrusion Detection: Detecting network attacks, usually by recognizing attack signatures extracted from earlier attack packets.

Intrusion Prevention: Protecting networks against attacks by taking some preemptive action such as access control, preventing the transmission of invalid IP addresses, and so forth.

Intrusion Tracking: The process of tracking an attack to its point of origin.

IP Traceback: The process of tracking the attack packets back to their source along the route they traveled.

ISP (Internet Service Provider): Refers to a company that provides access to the Internet and other related services (e.g., web hosting) to the public and other companies.

Network Intrusion: Broadly used to indicate stealing, destroying, or altering information, or obstructing information availability.

Network-Based Information System Model for Research

Jo-Mae B. Maris
Northern Arizona University, USA

INTRODUCTION

Cross-discipline research requires researchers to understand many concepts outside their own discipline. Computers are becoming pervasive throughout all disciplines, as evident by the December 2002 issue of *Communications of the ACM* featuring "Issues and Challenges in Ubiquitous Computing" (Lyytinen & Yoo, 2002). Researchers outside of computer network-related disciplines must account for the affects of network-based information systems on their research. This paper presents a model to aid researchers with the tasks of properly identifying the elements and affects of a network-based information system within their studies.

The complexity associated with network-based information systems may be seen by considering a study involving the effectiveness of an ERP for a mid-sized company. Such a study can become muddled by not recognizing the differences between the myriad of people, procedures, data, software, and hardware involved in the development, implementation, security, use, and support of an ERP system. If a researcher confuses network security limitations on users' accounts with ERP configuration limitations, then two important aspects of the information system being studied are obscured. One aspect is that a network must be secured so that only authorized users have access to their data. The other aspect concerns restrictions imposed by an ERP's design. Both aspects relate to the availability of data, but they come from different parts of the system. The two aspects should not be addressed as if both are attributable to the same source. Misidentifying network-based information system elements reflects negatively upon the legitimacy of an entire study.

BACKGROUND

Management information systems, applications systems development, and data communications each have contributed models that may be useful in categorizing network-based information system elements of a study. Kroenke (1981, p.25) offered a five-component model for planning business computer systems. Willis, Wilton, Brown, Reynolds, Lane Thomas, Carison, Hasan, Barnaby, Boutquin, Ablan, Harrison, Shlosberg, and Waters (1999, chap.1) discussed several client/server architectures for network-based applications. Deitel, Deitel, and Steinbuhler (2001, pp.600-620) presented a three-tier client/server architecture for network-based applications. The International Organization for Standardization (ISO) created the Open Systems Interconnection (OSI) Model (1994) for network communications. Zachman (2004) proposes a 30-cell matrix for managing an enterprise.

Kroenke's (1981, chap.2) five components are: people, procedures, data, software, and hardware. Procedures refer to the tasks that people perform. Data include a wide range of data from users' data to the data necessary for network configuration. Data form the bridge between procedures and software. Software consists of programs, scripts, utilities, and applications that provide the ordered lists of instructions that direct the operation of the hardware. The hardware is the equipment used by users, applications, and networks. Although Kroenke's five components are decades old, recent publications still cite the model including Kamel (2002), Pudyastuti, Mulyono, Fayakun and Sudarman (2000), Spencer and Johnston (2002, chap. 1), and Wall (2001).

The three-tiered model presented by Willis et al. (1999, pp.17-19) and Deitel et al. (2001, appendix B)

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

views a network-based application as consisting of a "client" tier, a middleware tier, and a "server" tier. The client tier contains applications that present processed data to the user and receives the user's data entries. The middleware processes data using business logic, and the server tier provides the database services. Another variation on the three-tiered model is found in Dean (2002, p.366). Dean's three-tiered model refers to client computers in networks. Her model consists of clients, middleware, and servers. In Dean's model, a client is a workstation on a network. The middleware provides access to applications on servers. Servers are attached to a network accessible by a client.

The ISO's OSI, as described by ISO (1994), is a seven-layer model used to separate the tasks of data communication within a network. The seven layers are physical, data link, network, transport, session, presentation, and application. The model describes services necessary for a message to travel from one open system to another open system.

The highest layer of the OSI model, applications, provides access to the OSI environment. The application layer provides services for other programs, operating system, or troubleshooting. For example, HTTP is an application layer utility. HTTP provides transfer services for Web browsers. The browser is not in the OSI application layer. The browser is above the OSI model.

"The presentation layer provides for common representation of the data transferred between application-entities" (ISO, 1994, clause 7.2.2.2). The services provided by the presentation layer include agreeing to encoding and encrypting schemes to be used for data being transferred. The presentation layer does not refer to formatted displays of a Web browser.

The remaining five layers of the OSI model pertain to contacting another node on a network, packaging, and addressing a message, sending the message, and assuring that the message arrives at its destination. The OSI model offers a framework for many vendors to provide products that work together in open systems. The OSI model does not encompass all of the components of a network-based information system.

Zachman's "Enterprise Architecture" (2004) consists of six rows and six columns which form 36 unique cells. The rows represent different levels of abstraction or development of an enterprise. The columns resolve who, what, where, when, why, and how. In order to use Zachman's model requires prior knowledge of the elements of network-based information systems, development, and operation.

Each of these models provides an answer to part of the puzzle for classifying the elements of a network-based information system. However, one must be familiar with the different types of personnel, procedures, data, software, and hardware to understand which are client, which are middleware, and which are server. One must be familiar with the different views of a system to determine which level of abstraction to use. If one delves further into the network's function, then the OSI model becomes important in understanding how a message is passed from a sender to a receiver. None of these models were intended to aid researchers outside of computing technologies areas to understand relationships among elements of a network-based information system.

MODEL

To help in understanding the different types of personnel, procedures, data, software, and hardware and how they work together, the following model is proposed.

Basic Network-based Information System Model

Let us begin with a three-tiered model. The top tier will represent the people who use the system and the people who benefit from the system's use. These people have procedures to follow in order to use the system or to receive the benefits. Also, the data representing information of interest to the users and beneficiaries would be in this top tier. For ease of reference this tier needs a name. Let us refer to the top tier as the specialty tier.

Next, let us have a middle tier that represents the applications that do the work the people represented in the specialty tier want performed. The middle tier we will call the application tier. In the application tier we would find a vast assortment of useful programs, such as Notepad, Word, Excel, StarOffice, SAP, Citrix, MAS 90, POSTman, SAS, RATS, Oracle, and countless others. These applications are above the

Table 1. Overview of network-based information system

Specialty Tier People, procedures, and data used to do work.
Application Tier Software used to do work. People, procedures, and data used to create, implement, and maintain software.
Infrastructure Tier People, procedures, system data, system software, and hardware necessary for the applications and network to operate satisfactorily.

OSI model and may receive services from the utilities in the application layer of the OSI model.

The bottom tier of our three-tiers will represent the workstations, operating systems, networking protocols, network devices, cabling, and all of the other hardware, people, procedures, data, and software necessary to make the network operate satisfactorily. Let us call this the infrastructure tier.

The model at this point appears as shown in Table 1. For some studies, this will provide a sufficient amount of detail.

Refined Network-based Information Systems Model

The model in Table 1 is very simplistic. It does not include organizational structures, inter-organizational connections, or customer-organization relationships that can exist in the specialty tier. Table 1 does not show the tiers of a network-based application, and it does not show the layers of communication in the network. Therefore, some studies may need a more detailed model that better defines the content of each tier.

When considering the details of the specialty tier, we will defer to the specialists doing the studies. Each study will organize the people, procedures, and data in the specialty tier as best fits the area being studied. For accounting and finance, generally accepted accounting principles (FASAB, 2004), Security and Exchange Commission filings and forms (SEC, 2004), and other financial standards and theories define the data. In management, organizational theory (AMR, 2003) gives guidance as to structures in which people work. Operations management (POMS, 2004) provides definitions of procedures used to produce goods and services. Marketing (AMA, 2004) has definitions for procurement procedures, data about goods and services, and relationships among people. Each functional area has its own rich resources for categorizing the elements represented by the specialty tier.

The three-tiered model of Deitel et al. (2001, appendix B) provides a meaningful classification scheme for the application tier. Thus, we can further classify applications within the application tier as to their place in three-tier architecture of presentation, functional logic, and data support.

Presentation applications are those that run on the local workstation and present data that have been processed. Examples of presentation applications are browsers, audio/video plug-ins, and client-side scripts. An SAP client is another example of a presentation application. Many standalone applications, such as NotePad, Word, PowerPoint, and StarOffice, are considered presentation applications.

Functional logic processes data from the data support sub-tier according to purpose specific rules. The functional logic then hands data generated to the presentation sub-tier. Examples of functional logic applications are Java server pages (JSP), Active Server Pages (ASP), and Common Gateway Interface (CGI) scripts. Enterprise resource planning (ERP) applications process data from a database and then hand processed data to client software on a workstation for display. ERP applications are examples of functional logic applications.

Data support applications manage data. Databases are the most common example. However, applications that manage flat files containing data are also data support applications. In general, a program that adds, deletes, or modifies records is a data support application.

The OSI model refines much of the infrastructure tier. As presented in Dean (2002, chap.2), the OSI model describes communications of messages, but it does not include operating systems, systems utilities, personnel, system data above its applications layer, and procedures necessary for installing, configuring, securing, and maintaining the devices and applications represented by the infrastructure tier. If a study involves significant detail in the infrastructure tier, then the research team should include an individual familiar with network technology and management.

Table 2. Refined network-based information system

Specialty Tier Models, standards, or theories specific to information use being studied
Application Tier • Presentation applications present processed data • Functional logic applications process data according to rules usually known to the researchers • Data support applications manage data
Infrastructure Tier • People, procedures, and data necessary to install, configure, secure, and maintain devices, applications, and services in system • System applications outside the OSI Model • OSI Model o *Applications Layer*: communications protocols, such as HTTP, FTP, and RPC o *Presentation Layer*: data preparation, such as encryption o *Session Layer*: protocols for connecting sender and receiver o *Transport Layer*: initial subdividing of data into segments, error checking, and sequencing segments o *Network Layer*: packaging segments into packets with logical addressing and routing packets o *Data Link Layer*: packaging packets into frames for specific type of network with physical addressing o *Physical Layer*: signal representing frames, media and devices carrying signals

Table 2 summarizes the refined model for categorizing elements of a network-based information system.

USING NETWORK-BASE INFORMATION SYSTEM MODEL

Let us consider some examples of research using this model. First imagine a study investigating the effects of using an ERP's accounting features upon the financial performance of mid-size firms. For the specialty tier, accountants conducting the study decide classifications for users, users' procedures, and data structures. In the application tier would be several products related to the ERP. In the presentation sub-tier would be the ERP's client that runs on the workstations used by participants in the study. The ERP business rules modules would be in the functional logic sub-tier. The database used by the ERP would be in the data support sub-tier. In the infrastructure tier, we would find the hardware and system software necessary to implement and support the network, ERP, and the database.

Now let us look at a few events that might occur during the study. A user may not be able to log into the network. This event should be attributed to the infrastructure tier, and not presented as a deficiency of the ERP. On the other hand, if the accountants found that the calculation of a ratio was incorrect, this problem belongs to the application tier, in particular the functional logic. Another problem that might arise is users entering incorrect data. The entering of incorrect data by users would be a specialty tier problem. By properly classifying the elements in the study related to the network-based information system, the researchers may find a conflict between the specialty tier definition of data and the application tier definition of data. Since the elements of the system are properly defined, the data definition conflict will be more obvious and more easily substantiated.

Now let us consider another possible study. Suppose marketing researchers are investigating the effectiveness of Web pages in selling XYZ product. The marketing researchers would decide on classifications of people involved in the use of the Web site, procedures used by the people, and structure of data involved in the use of the Web site. All of these definitions would be in the specialty tier. The browser used to display Web pages would be in the presentation sub-tier of the application tier. The server-side script used to produce Web pages and apply business rules would be in the functional logic sub-tier of the application tier. The database management system used to manage data used in the Web site would be in the data support sub-tier of the application tier. The Web server executing server-side scripts and serving Web pages would be in the infrastructure tier.

By properly classifying elements of the study, the marketing researchers would be able to better assess the marketing specific effects. A few events that might have occurred during the study include failure of a user to read a Web page, an error in computing quantity discount, and an aborted ending of a session due to a faulty connection. The failure of a user to read a Web page would be a specialty tier error. The incorrect computation of the quantity discount would be an application tier error in the functional logic. The aborted session would be an infrastructure tier error. If the researchers are primarily interested in the user's reaction to the Web pages, they may have delimited the study so that they could ignore the infrastructure tier error.

Differentiating between the tiers can be problematic. For example, Access can be seen as encompassing all three application sub-tiers. The forms, reports, and Web pages generated by Access are usually considered to belong to the presentation sub-tier. Modules written by functional area developers belong to the functional logic sub-tier. The tables, queries, and system modules are in the data support sub-tier. In the marketing study about selling XYZ product on the Web, the researchers may need to distinguish a query that gets catalog items stored in an Access database, from an ASP page that applies customer specific preferences, from a Web page that displays the resulting customer-specific catalog selections. In this case, Access would be in the data support sub-tier, the ASP page would be in the functional logic sub-tier, and the Web page would be in the presentation sub-tier.

In some studies, even applications that we normally think of as presentation sub-tier applications may have elements spread across the three tiers of the model. For example, a business communications study may be interested in formatting errors, grammar errors, file corruption, and typographical errors. In this study, typographical errors could be due to specialty tier error or infrastructure tier problems. If the user makes a mistake typing a character on a familiar QWERTY keyboard, then the error would be attributed to the specialty tier. On the other hand, if the user were given an unfamiliar DVORAK keyboard, then the error could be attributed to the Infrastructure Tier. In neither case should the errors be attributed to the application tier. Within an application, Word incorrectly reformatting may be classified as a presentation feature's error. Grammar errors undetected by the grammar checker could be attributed to the functional sub-tier, and the corruption of a file by the save operation would be a data support error.

As just shown, classifying the elements and events may vary according to each study's purpose. However, the classification should reflect appropriate network-based information system relationships. In differentiating the application tier elements, a helpful tactic is to identify which elements present data, which elements perform functional area-specific processing, and which elements manage data. Once the elements are identified, then the events associated with those elements are more easily attributed. Properly classifying information system elements and ascribing the events make a study more reliable.

CONCLUSION

In this paper, we have seen that the elements and events of a study involving a network-based information system may be classified to reduce confusion. The appropriate classification of information system elements and attribution of events within a study should lead to more reliable results.

REFERENCES

Academy of Management Review (AMR, 2003). ARM Home Page. Retrieved May 19, 2004, from *http://www.aom.pace.edu/amr/*

American Marketing Association (AMA, 2004). American Marketing Association: the source. Retrieved May 19, 2004, from *http://www.marketingpower.com/*

Dean, T. (2002). *Network+ Guide to Networks, 2nd ed.* Boston, Massachusetts: Course Technology.

Deitel, H., Deitel, P. & Steinbuhler, K. (2001). *E-business and e-commerce for managers*. Upper Saddle River, New Jersey: Prentice Hall.

Federal Accounting Standards Advisory Board (FASAB, 2004). Generally Accepted Accounting Principles. Retrieved on May 18, 2004, from *http://www.fasab.gov/accepted.html*

International Organization for Standardization (ISO, 1994) (Obsoletes 1984). ISO/IEC 7498 The Basic Model, part 1: Information Processing Systems – OSI Reference Model – The Basic Model. Retrieved on May 6, 2004, from *http://www.acm.org/sigcomm/standards/iso_stds/OSI_MODEL/ISO_IEC_7498-1.TXT*

Kamel, S. (2002). The Use of DSS/EIS for Sustainable Development in Developing Nations, *Proceedings of InSite 2002*. Retrieved on May 5, 2004, from *http://ecommerce.lebow.drexel.edu/eli/2002Proceedings/papers/Kamel239useds.pdf*

Kroenke, D. (1981). *Business computer systems: An introduction*. Santa Cruz, California: Mitchell Publishing Inc.

Lyytinen, K. & Yoo Y. (2002). Issues and Challenges in Ubiquitous Computing. *Communications of the ACM,* (12), 63-65.

Production Operations Management Society (POMS, 2004). POMS Home. Retrieved on May 19, 2004, from *http://www.poms.org/*

Pudyastuti, K., Mulyono, A., Fayakun, F. & Sudarman, S. (2000). Implementation of the management information system in Pertamina – Geothermal division. In *Proceedings of World Geothermal Congress*. Retrieved on May 14, 2004, from *http://www.geothermie.de/egec-geothernet/ci_prof/asia/indonesia/0618.pdf*

Securities and Exchange Commission (2004). Filings and Forms (EDGAR). Retrieved on May 18, 2004, from *http://www.sec.gov/edgar.shtml*

Spencer, H. & Johnston, R. (2002). *Technology best practices*. Indianapolis, IN: John Wiley and Sons.

Wall, P. (2001). Centralized versus Decentralized Information Systems in Organizations. Retrieved May 14, 2004, from *http://emhain.wit.ie/~pwall/CvD.htm*

Willis, T., Wilton, P., Brown, M., Reynolds, M, Lane Thomas, M., Carison, C., Hasan, J., Barnaby, T., Boutquin, P., Ablan, J., Harrison, R., Shlosberg, D., & Waters, T. (1999). *Professional VB6 Web Programming*. West Sussex, England: Wrox Press Ltd.

Zachman, J. (2004). Concepts of the Framework for Enterprise Architecture: Background, Description and Utility. Retrieved March 17, 2004, from *http://members.ozemail.com.au/~visible/papers/zachman3.htm*

KEY TERMS

Application: An application is a program, script, or other collection of instructions that direct the operation of a processor. This is a wide definition of "application." It does not distinguish Web-based software from standalone software. Nor does this definition distinguish system software from goal specific software.

Client: A client is a computer, other device, or application that receives services from a server.

Device: A device is a piece of equipment used in a network. Devices include, but are not limited to, workstations, servers, data storage equipment, printers, routers, switches, hubs, machinery or appliances with network adapters, and punch-down panels.

Network: A network consists of two or more devices with processors functioning in such a way that the devices can communicate and share resources.

Operator: An operator is a person who tends to the workings of network equipment.

Record: A record is composed of fields that contain facts about something, such as an item sold. Records are stored in files.

Server: A server is a computer or application that provides services to a client.

User: A user is a person who operates a workstation for one's own benefit or for the benefit of one's customer.

Workstation: A workstation is a computer that performs tasks for an individual.

A New Block Data Hiding Method for the Binary Image

Jeanne Chen
HungKuang University, Taiwan

Tung-Shou Chen
National Taichung Institute of Technology, Taiwan

Meng-Wen Cheng
National Taichung Institute of Technology, Taiwan

INTRODUCTION

Great advancements in Web technology have resulted in increase activities on the Internet. Users from all walks of life — e-commerce traders, professionals and ordinary users — have become very dependent on the Internet for all sorts of data transfers, be it important data transactions or friendly exchanges. Therefore, data security measures on the Internet are very essential and important. Also, steganography plays a very important role for protecting the huge amount of data that pass through the internet daily.

Steganography (Artz, 2001; Chen, Chen & Chen, 2004, Qi, Snyder, & Sander, 2002) is hiding data into a host image or document as a way to provide protection by keeping the data secure and invisible to the human eyes. One popular technique is to hide data in the least significant bit (LSB) (Celik, Sharma, Tekalp, & Saber, 2002; Tseng, Chen, & Pan, 2002). Each pixel in a gray image takes up eight bits representation and any changes to its last three LSBs are less likely to be detected by the human visual system. However, an image in LSB hiding is not robust to attacks.

Other hiding techniques involve robust hiding (Lu, Kot, & Cheng, 2003), bit-plane slicing (Noda, Spaulding, Shirazi, Niimi, & Kawaguchi, 2002), hiding in compressed bitstreams (Sencar, Akansu, & Ramkumar, 2002) and more. Some applications require more data to be hidden but must not have visually detectable distortions. An example would be the medical records. More details on high capacity hiding could be found in Moulin and Mihcak (2002), Candan and Jayant (2001), Wang and Ji (2001), Rajendra Acharya, Acharya, Subbanna Bhat, and Niranjan (2001), and Kundur (2000).

Although much research had been done for data hiding, very little exists for hiding in binary images. The binary (or black and white) image is common and often appears as a cartoon in newspapers and magazines. Most are easy preys to piracy. Hiding is difficult for the binary image, since each of its black or white pixels requires only one bit representation. Any bit manipulation will reveal hiding activities, and the image is easily distorted. Indeed, the block data hiding method (BDHM) proposed in this paper is concentrated largely for hiding in binary images.

RELATED WORK: A NOVEL HIDING METHOD

Pan, Wu, and Wu (2001) partitioned an image into blocks where each block would be repartition into four overlapping sub-blocks. Next, all the white pixels in the sub-blocks will have a number assigned to each of them. Based on these numbers, the characteristic values of the sub-blocks were calculated and used to determine the suitable sub-blocks for hiding such that the hidden data will show as uniformly distributed on the block and have no visible distortions. Data will only be hidden in the center pixel. The bit for the center pixel in the selected sub-block will be toggled from white to black as hiding a bit.

For example, a 512×512 image which was partitioned into 16384 units of 4×4 blocks as in Figure 1(a). Each block was further repartitioned into four overlapping 3×3 sub-blocks as in Figure 1(b). From

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

the sub-blocks, it can then be easily determined that the possible sub-blocks for hiding are (1), (2) and (11). Also, after hiding the characteristic values should not be altered.

THE PROPOSED BLOCK DATA HIDING METHOD (BDHM)

A binary image is actually made up of black and white pixels. There is only one bit representation for each pixel. Each pixel is either 0 for black or 1 for white. There are two types of binary image: one is the simple black-white binary image, and the other is the complex binary image such as the natural or halftone images.

The Block Data Hiding Method (BDHM) proposed in this paper involves data hiding and data retrieving for the black-white binary image. First, the original image is partitioned into blocks. Next, characteristic values for individual blocks will be calculated and important data hidden in the blocks based on these values. The image with hidden data is called a stego-image. The data retrieval process is similar to data hiding with the exception that data will be retrieved. Figure 2 illustrates the flow for data hiding and retrieving.

Partitioning into Blocks and Sub-Blocks

First, the simple binary image will be partitioned into $N \times N$ sized blocks. Next, each $N \times N$ sized block will be repartitioned into four overlapping $n \times n$ sized blocks. Suppose $N \times N$ is a 4×4 sized block, then the overlapping sub-blocks will be 3×3 as shown in Figure 3. Pixels in columns two and three of sub-block (1) overlap pixels in columns one and two of (2). Pixels for sub-blocks (3) and (4) also overlap in this respect, too.

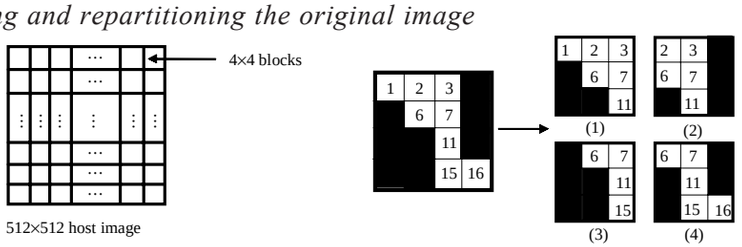
Calculating Characteristic Values

The characteristic values for the blocks are calculated based on the number of white pixels (each bit value is 1). Let R_j be the characteristic value for block j where j=1, 2, 3, …, M. M is the total number of partitioned blocks. Let r_i be the number of white pixels in sub-block i where i=1, 2, 3, 4 and T_j be the total number of sub-blocks with the same least number of white pixels. Then $R_j=r_i$ where r_i is the least number of white pixels. As illustrated in the example in Figure 4, r_i={3, 3, 4, 3} for sub-blocks (1), (2), (3) and (4), respectively. Since three sub-blocks; (1), (2), and (4) have the least number of white pixels, T_j=3. The least number of white pixels in the sub-blocks is 3; therefore R_j=3.

Hiding the Data

After characteristic values had been calculated for every block, the blocks will be sorted by ascending order of R_j. No data must be hidden in B_j if it contains exclusively black or white pixels. If a white pixel is hidden in an all black block (R_j=9) or a black pixel in an all white block (R_j=0), the contrast is too noticeable to the naked eyes. Therefore to avoid hiding in exclusively black or white blocks, a rule is set to allow hiding only in blocks with 3f"R_jf"6. The white and black pixels are usually distributed uniformly in the blocks with R_j within this range. Further information would also be needed on the distribution of the black and white pixels in each block before starting the hiding process. Figure 5 shows different distributions for two different blocks. Although the characteristic value R_j in Figure 5(a) is smaller than Figure 5(b), the distribution of the black and white pixels is more uniform in Figure 5(b). Data hidden in Figure 5(a) would stand out clearly rather than in Figure 5(b).

Figure 1. Partitioning and repartitioning the original image



(a) Partitioning into 4×4 blocks (b) Repartitioning into 3×3 sub-blocks

Figure 2. Flow for data hiding and retrieving(1)

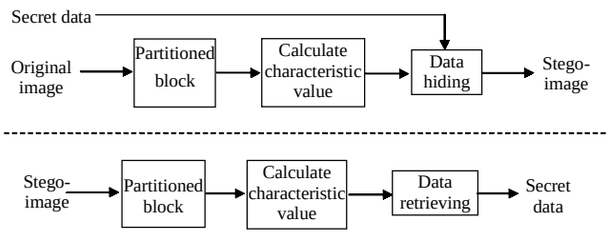


Figure 3. The 4×4 block repartitioned into 3×3 sub-blocks(2)

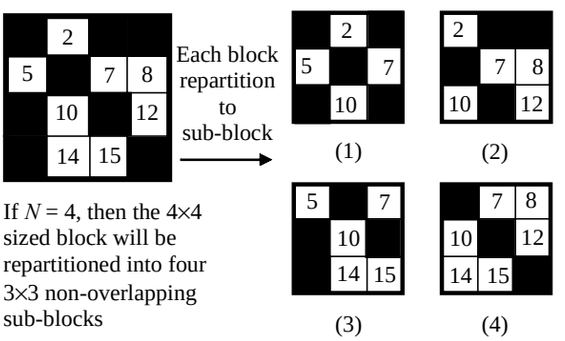

Each block repartition to sub-block
If N = 4, then the 4×4 sized block will be repartitioned into four 3×3 non-overlapping sub-blocks
(1)
(2)
(3)
(4)

Figure 4. Calculating characteristic values for the sub-blocks

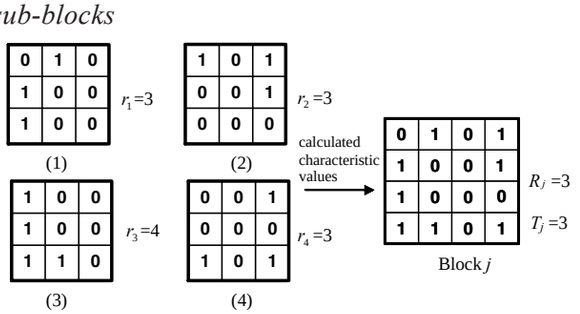
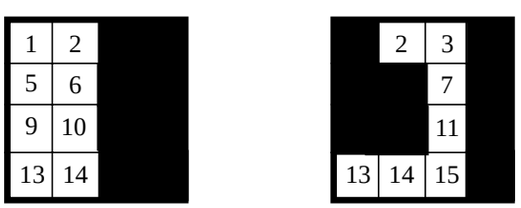


Figure 5. Distribution of the black and white pixels



(a) Concentrated distribution (b) Uniform distribution

To find the block to hide, let W_j be the most number of neighbor pixels with the same pixel value in block B_j. A bigger W_j implies that B_j has a dense distribution; otherwise it is uniform. Then a weighted value F_j can be calculated by adding W_j and R_j for only when $3 \leqq R_j \leqq 6$. A smaller F_j implies that block B_j is suitable for hiding.

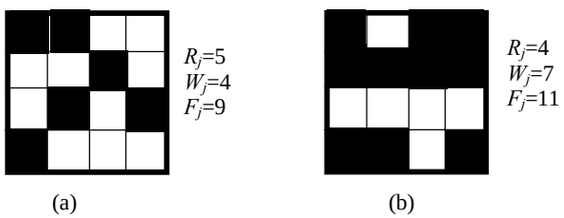
All the F_j will be sorted in ascending order. Data will be hidden in the ascending sequence. A smaller R_j means that block B_j has less white pixels. Meanwhile, a smaller F_j means that the distribution of pixels is more uniform. Data hidden in the block with a smaller F_j will not be easy to detect. Figure 6(a) illustrates an example with R_j=5, W_j=4 and $F_j=W_j+R_j$=9. Figure 6(b) shows R_j=4, W_j=7 and $F_j=W_j+R_j$=11. The weighted value F_j in Figure 6(a) is smaller than in Figure 6(b). The black and white pixels in Figure 6(a) are more uniformly distributed than in Figure 6(b). Therefore, Figure 6(a) has a better hiding block than Figure 6(b).

After locating the block for hiding, T_j will be used to decide on the hiding location. Suppose $En(D)$ is the hidden data, and then the simple rule for hiding is as follows.

$$En(D) = \begin{cases} 1, & \text{make T = 1, 3 else} \\ 0, & \text{make T = 2, 4.} \end{cases}$$

Therefore, to hide in a pixel that has a value 1 and T_j is 2 or 4; T_j must be modified to 1 or 3. Conversely, if the pixel has a value 0 and T_j is 1 or 3; T_j must be modified to 2 or 4. Care must be taken so that the characteristic value R_j will not be modified when data is hidden into the block. Figure 7 illustrates the hiding process. In the example, the 4×4 block B_j is partitioned into four overlapping 3×3 sub-blocks. Then r_1=3, r_2=3, r_3=4 and r_4=3 for the respective sub-blocks. The characteristic value R_j is 3, and T_j=3. The pixel where data is to be hidden has a value 1 and T_j=3. Let the data to hide be 0. Therefore, T_j must be changed to 2 or 4. The third sub-block is the one most suitable for hiding and the bit on the lower-left corner is toggled from 1 to 0, while T_j switches from 3 to 4. Only pixels on the circumference of the sub-block will be chosen for hiding.

Figure 6. Calculating the W_j and F_j values.



A flexible amount of data can be hidden depending on the number of available blocks that have R_j in the scope between 3 and 6. More could be hidden if more available blocks are located. The proposed method is suitable for a complex natural image that contains uniformly mixed black and white pixels. Pan et al.'s (2001) method is only suitable for the simple black and white binary image. With their method, less data can be hidden since data are hidden in the overlapped center pixels of the sub-blocks. The stego-image in their experiment showed artifacts. In the proposed method data are hidden on the circumference and the R_j cannot be changed from the hiding process. These combinations will not allow the hidden data to be easily detected. The flow for the proposed hiding method is illustrated in Figure 8.

Retrieving the Hidden Data

The steps for retrieving the hidden data are a mimic of the hiding process with the exception that the hidden data is being retrieved. First, the stego-image is partitioned into $N \times N$ blocks and then into overlapping sub-blocks. Values like r'_j, R'_j, T'_j, W'_j and F'_j similar to the ones in the hiding process (r_j, R_j, T_j, W_j and F_j respectively), will be calculated for each block. T'_j will be used to determine which hidden data is to be retrieved. The data to be retrieved should be in one of the pixels on the circumference of block B'_j. Supposing T'_j is 1 or 3, then the hidden data to retrieve is 1; otherwise, if T'_j is 2 or 4, the hidden data to retrieve is 0. This completes the data retrieval process.

EXPERIMENTAL ANALYSIS AND DISCUSSION

In the experiments, four 256×256 binary images: "owl", "doll", "gorilla", and "vegetables" were used. Judgments will be made by raw eye visual comparison between the original image and the stego-image for any visually detectable changes.

Figure 9(a) illustrates an "owl" binary image. Precaution is taken not to hide in the exclusively black

Figure 7. Data hiding process

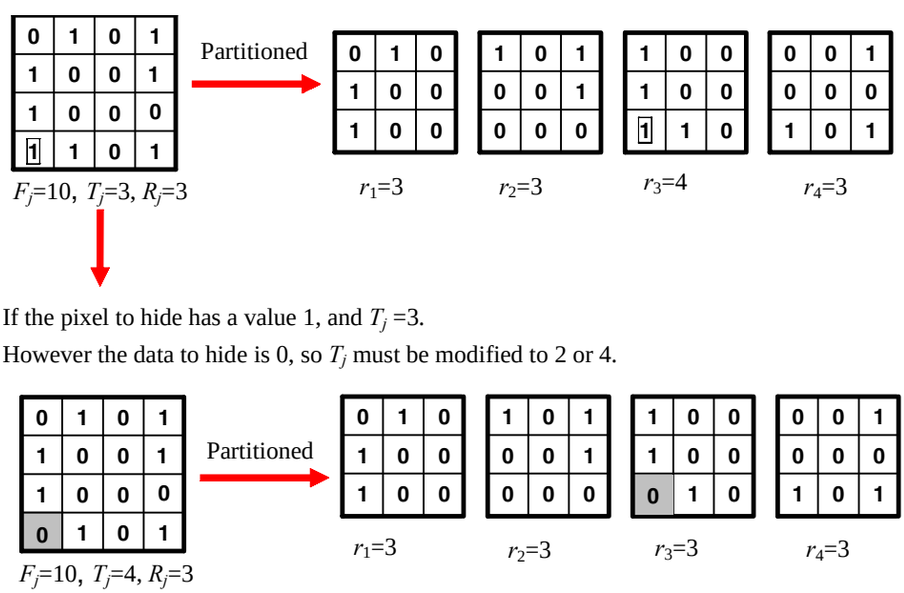
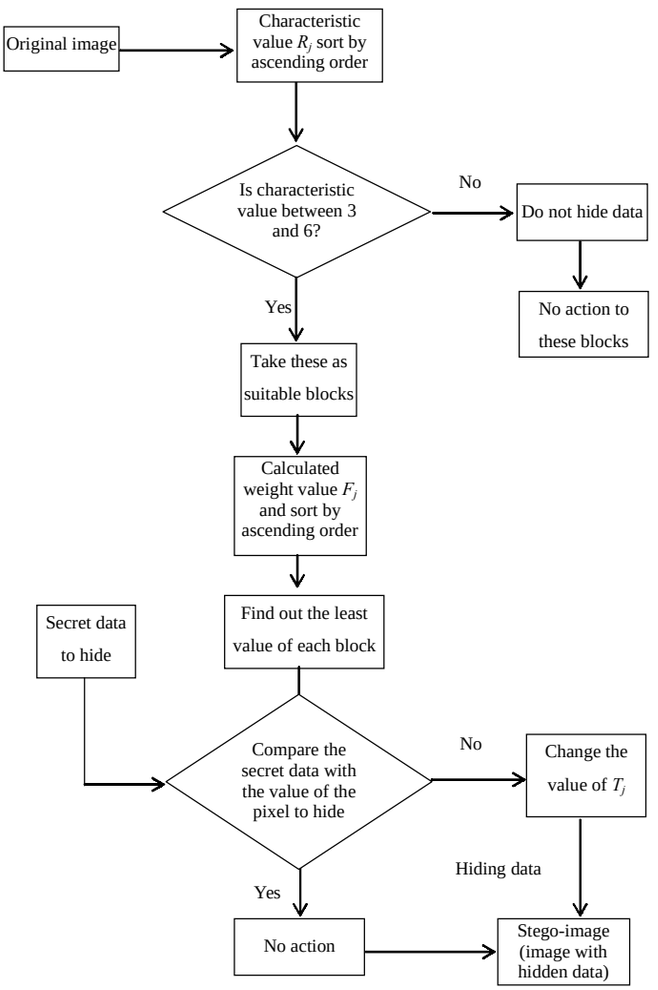


Figure 8. Data hiding flowchart


Original image
Characteristic value R_j sort by ascending order
Is characteristic value between 3 and 6?
No
Do not hide data
No action to these blocks
Yes
Take these as suitable blocks
Calculated weight value F_j and sort by ascending order
Secret data to hide
Find out the least value of each block
Compare the secret data with the value of the pixel to hide
No
Change the value of T_j
Hiding data
Yes
No action
Stego-image (image with hidden data)

or white blocks. Projected calculations showed 335 blocks can be used for hiding data. Therefore, a maximum of 335 bits were hidden in Figure 9(b). By visual comparison, there appears to be some tiny differences on the edges of the leaf and the head of the owl; these are the areas with hidden data.

Figure 10(a) illustrates the "doll" binary image. A maximum of 208 bits were hidden in Figure 10(b). The areas around the doll's clothes show differences between the original image and the stego-image.

Figure 11(a) illustrates the "gorilla" binary image. A maximum of 956 bits were hidden in Figure 11(b). Since the pixels in "gorilla" are very uniformly distributed, the hidden data cannot be easily detected.

Similarly, in Figure 12(b), 315 bits were hidden. Figure 12(b) shows slight differences in the eggplant's lower left side. These are the areas of hidden data.

From the experimental results, the quality of the stego-images is comparable to the original images. Furthermore, the stego-image in Figure 11(b) showed similar quality to the original. This is because the black and white pixels on the image are more uniformly distributed.

Figure 9. The "owl" binary image: 335 bits hidden

(a) The original 256×256 image

(b) Image after hiding

Figure 10. The "doll" binary image: 208 bits hidden

(a) The original 256×256 image

(b) Image after hiding

Figure 11. The "gorilla" binary image: 956 bits hidden

(a) The original 256×256 image

(b) Image after hiding

Figure 12. The "vegetable" binary image: 315 bits hidden

(a) The original 256×256 image

(b) Image after hiding

CONCLUSION

As illustrated in Figure 11(b), this method works well for images with uniformly distributed black and white pixels. In the other experimental samples although some differences appear in the stego-images—they do not affect the appearance of the originals. This is a simple method that can be useful for providing basic authentication for the simple binary images. These kinds of images appear commonly on newspapers, comics, and more. Due to their popular use, they are often preys of piracies. In the experiments, only arbitrary bits were hidden. Instead, watermarks or logos could be hidden and to be retrieved later on for use in authenticating an image. More details on watermarks and improved security could be found in Kirovski and Petitcolas (2003).

Future work could be to improve the security of the hidden data and to increase the capacity for hiding. Tseng et al. (2002) used a secret key with a weighted matrix to protect the hidden data. Also, they were able to manipulate bits in a block to allow hiding up to 2 bits. The DBHM discussed in this paper hides in one bit per block. However data is hidden in the pixel on the circumference of a block. There are 16 pixels making up the circumference in a 4×4 block. Therefore, in theory a maximum of 16 bits could be hidden in a 4×4 block, which is eight times that of Tseng et al.'s method. In DBHM the hiding method is easy to hack and the hidden data extracted and replaced with a faked one. In future work, the watermark or logo could be encrypted before being hidden. A potential hacker would than have a tough time trying to decrypt the hidden data.

REFERENCES

Alturki, F. & Mersereau, R. (2001). A novel approach for increasing security and data embedding capacity in images for data hiding applications. *Proceedings of the International Conference on Information Technology: Coding and Computing*, (pp. 228-233).

Artz, D. (2001). Digital steganography: Hiding data within data. *IEEE Internet Computing*, *5*(3), 75-80.

Candan, C. & Jayant, N. (2001). A new interpretation of data hiding capacity. *Proceedings on the 2001 IEEE International Confererence on Acoustics, Speech and Signal Processing (ICASSP '01)*, May 7-11, (Vol. 3, pp. 1993-1996).

Celik, M.U., Sharma, G., Tekalp, A.M., & Saber, E. (2002). Reversible data hiding. *Proceedings on the 2002 International Conference on Image Processing*, (Vol. 2, pp. II-157-II-160).

Chen, T.S., Chen, J., & Chen, J.G. (2004). A simple and efficient watermark technique based on JPEG2000 Codec. *ACM Multimedia Systems Journal*, 16-26.

Kirovski, D. & Petitcolas, F.A.P. (2003). Blind pattern matching attack on watermarking systems. *IEEE Transactions on Signal Processing*, *51*(4), 1045-1053.

Lu, H., Kot, A.C., & Cheng, J. (2003). Secure data hiding in binary document images for authentication. *Proceedings of the 2003 International Symposium on Circuits and Systems (ISCAS '03)*, (Vol. 3, pp. III-806-III-809).

Moulin, P. & Mihcak, M.K. (2002). A framework for evaluating the data-hiding capacity of image sources. *IEEE Transactions on Image Processing*, *11*(9), 1029-1042.

Noda, H., Spaulding, J., Shirazi, M.N., Niimi, M., & Kawaguchi, E. (2002). Application of bit-plane decomposition steganography to wavelet encoded images. *Proceedings of the 2002 International Conference on Image Processing*, (pp. II-909-II-912).

Pan, G., Wu, Y.J., & Wu, Z.H. (2001). A novel data hiding method for two-color images. *Lecture Notes in Computer Science Information and Communications Security,* (pp. 261-270).

Qi, H., Snyder, W.E., & Sander, W.A (2002). Blind consistency-based steganography for information hiding in digital media. *Proceedings of the 2002 IEEE International Conference on Multimedia and Expo*, (Vol. 1, pp. 585-588).

Rajendra, U., Acharya, D., Subbanna Bhat, P. & Niranjan, U.C. (2001). Compact storage of medical images with patient information. *IEEE Transactions on Information Technology in Biomedicine*, *5*(4), 320-323.

Sencar, H.T., Akansu, A.N., & Ramkumar, M. (2002). Improvements on data hiding for lossy compression. *Proceedigns on IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, (Vol. 4, pp. IV-3449-IV-3452).

Tseng, Y.C., Chen, Y.Y., & Pan, H.K. (2002). A secure data hiding scheme for binary images. *IEEE Trans. on Communications*, *50*(8), 1227-31.

Tseng, Y.C. & Pan, H.K. (2002). Data hiding in 2-color images. *Computers, IEEE Transactions on*, *51*(7), 873-878.

Wang, J. & Ji, L. (2001). A region and data hiding based error concealment scheme for images. *IEEE Transactions on Consumer Electronics*, *47*(2), 257-262.

KEY TERMS

Binary Image: An image made up of black and white pixels with values of 0s or 1s.

Block Data Hiding Method (BDHM): In BDHM, an image will be partitioned into blocks and sub-blocks. Then based on the characteristic values of these sub-blocks, the most suitable sub-block will be chosen for hiding. Data hidden in the block will not be visually easy to detect and must not modify the original characteristic value of the block.

Data Hiding: Important data being embedded into a host image.

Encrypting: Data that is scrambled such that it appears meaningless to an average user. An authorized user can later restore the data back to its original.

Partition: An image that is divided into blocks for processing.

Steganography: To hide important data into a host image such that it is not easy to be detected.

Stego-Image: An image that has important data hidden within.

Tamper: Making alterations to an image with unfriendly intent.

Objective Measurement of Perceived QoS for Homogeneous MPEG–4 Video Content

Harilaos Koumaras
University of Athens, Greece

Drakoulis Martakos
National and Kapodistrian University of Athens, Greece

Anastasios Kourtis
Institute of Informatics and Telecommunications NCSR Demokritos, Greece

INTRODUCTION

Multimedia applications over 3G and 4G (third and fourth generation) networks will be based on digital encoding techniques (e.g., MPEG-4) that achieve high compression ratios by exploiting the spatial and temporal redundancy in video sequences. However, digital encoding causes image artifacts, which result in perceived-quality degradation. Due to the fact that the parameters with strong influence on the video quality are normally those set at the encoder (most importantly, the bit rate and resolution), the issue of user satisfaction in correlation with the encoding parameters has been raised (MPEG Test, 1999).

One of the 3G-4G visions is to provide audiovisual (AV) content at different qualities and price levels. There are many approaches to this issue, one being the perceived quality of service (PQoS) concept. The evaluation of the PQoS for multimedia and audiovisual content will provide a user with a range of potential choices, covering the possibilities of low, medium, or high quality levels. Moreover the PQoS evaluation gives the service provider and network operator the capability to minimize the storage and network resources by allocating only the resources that are sufficient to maintain a specific level of user satisfaction.

This paper presents an objective PQoS evaluation method for MPEG-4-video-encoded sources based on a single metric experimentally derived from the spatial and temporal (S-T) activity level within a given MPEG-4 video.

Toward this, a quality meter tool was used (Lauterjung, 1998), providing objective PQoS results for each frame within a video clip. The graphical representation of these results vs. time demonstrated the instant PQoS of each frame within the video clip, besides indicating the mean PQoS (MPQoS) of the entire video (for the whole clip duration). The results of these experiments were used to draw up experimental curves of the MPQoS as a function of the encoding parameters (i.e., bit rate). The same procedure was applied for a set of homogeneous video sequences, each one representing a specific S-T activity level.

Furthermore, this paper shows that the experimental MPQoS vs. bit-rate curves can be successfully approximated by a group of exponential functions, which confines the QoS characteristics of each individual video test sequence to three parameters. Showing the interconnection of these parameters, it is deduced that the experimental measurement of just one of them, for a given short video clip, is sufficient for the determination of the other two. Thus, the MPQoS is exploited as a criterion for preencoding decisions concerning the encoding parameters that satisfy a certain PQoS in respect to a given S-T activity level of a video signal.

BACKGROUND

Over the last years, emphasis has been put on developing methods and techniques for evaluating the perceived quality of video content. These meth-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

ods are mainly categorized into two classes: the subjective and objective ones.

The subjective test methods involve an audience of people who watch a video sequence and score its quality as perceived by them under specific and controlled watching conditions. The mean opinion score (MOS) is regarded as the most reliable method of quality measurement and has been applied on the most known subjective techniques: the single-stimulus continue quality evaluation (SSCQE) and the double-stimulus continue quality evaluation (DSCQE) (Alpert & Contin, 1997; ITU-R, 1996; Pereira & Alpert, 1997). However the MOS method is inconvenient due to the fact that the preparation and execution of subjective tests is costly and time consuming.

For this reason, a lot of effort has recently been focused on developing cheaper, faster, and easier applicable objective evaluation methods. These techniques successfully emulate the subjective quality-assessment results based on criteria and metrics that can be measured objectively. The objective methods are classified according to the availability of the original video signal, which is considered to be of high quality.

The majority of the proposed objective methods in the literature require the undistorted source video sequence as a reference entity in the quality-evaluation process, and due to this, they are characterized as full-reference methods (Tan & Ghanbari, 2000; Wolf & Pinson, 1999). These methods are based on an error-sensitivity framework with the most widely used metrics: the peak-signal-to-noise ratio (PSNR) and the mean square error (MSE).

$$\text{PSNR} = 10\log_{10}\frac{L^2}{\text{MSE}} \quad (1)$$

where L denotes the dynamic range of pixel values (equal to 255 for 8 bits/pixel monotonic signal).

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \quad (2)$$

where N denotes the number of pixels, and x_i/y_i is the i^{th} pixel in the original and distorted signal.

However, these overused metrics have seriously been criticized as not providing reliable measurements of the perceived quality (Wang, Bovik, & Lu, 2002). For this reason, a lot of effort has been focused on developing assessment methods that emulate characteristics of the human visual system (HVS) (Bradley, 1999; Daly, 1992; Lai & Kuo, 2000; Watson, Hu, & McGowan, 2001) using contrast-sensitivity functions (CSFs), channel decomposition, error normalization, weighting, and Minkowski error pooling for combining the error measurements into a single perceived-quality estimation. An analytical description of the framework that these methods use can be found in Wang, Sheikh, and Bovik (2003).

However, it has been reported (VQEG, 2000; Wang et al., 2002) that these complicated methods do not provide more reliable results than the simple mathematical measures (such as PSNR). Due to this, some new full-reference metrics that are based on the video structural distortion and not on error measurement have been proposed (Wang, Bovik, Sheikh, & Simoncelli, 2004; Wang, Lu, & Bovik, 2004).

On the other hand, the fact that these methods require the original video signal as reference deprives their use in commercial video-service applications where the initial undistorted clips are not accessible. Moreover, even if the reference clip is available, the synchronization predicaments between the undistorted and the distorted signal (which may have experienced frame loss) make the implementation of the full-reference methods difficult.

Due to these reasons, recent research has been focused on developing methods that can evaluate the PQoS based on metrics that use only some extracted features from the original signal (reduced-reference methods) (Guawan & Ghanbari, 2003) or do not require any reference video signal (no-reference methods) (Lauterjung, 1998; Lu, Wang, Bovik, & Kouloheris, 2002).

A software implementation that is representative of this nonreference objective evaluation class is the quality meter software (QMS) that was used for the needs of this paper (Lauterjung, 1998). The QMS tool measures objectively the instant PQoS level (on a scale from 1 to 100) of digital video clips. The metrics used by the QMS are vectors, which contain information about the averaged luminance differences of adjacent pixel pairs that are located across and on both sides of adjacent DCT-block (8x8

pixels) borders. At these pixel pairs, the luminance discontinuities are increased by the encoding process following a specific pattern, in contrast with the rest of the pixel pairs of the frame.

The validity of the specific QMS has been tested (Lauterjung, 1998) by comparing quality-evaluation results derived from the QMS to corresponding subjective quality-assessment results, which were deduced by an SSCQE subjective test procedure. This comparison showed that the QMS emulates successfully the corresponding subjective quality-assessment results.

Figure 1 depicts an example measurement of the instant PQoS derived from the QMS for the clip "Mobile & Calendar," which was encoded using the MPEG-4 standard (simple profile) at 800 Kbps (constant bit rate) with common intermediate format (CIF) resolution at 25 frames per second (fps). The instant PQoS vs. the time curve (where time is represented by the frame sequence) varies according to the S-T activity of each frame. For frames with high complexity, the instant PQoS level drops, while for frames with low S-T activity, the instant PQoS is higher.

Such instant PQoS vs. time curves derived by the QMS tool can be used to categorize a short video clip according to its content. Introducing the concept of the MPQoS, the average PQoS of the entire video sequence over the whole duration of a short clip can be defined as follows:

$$\text{MPQoS} = \frac{\sum_{i=1}^{N} \text{Instant PQoS}_i}{N} \quad (3)$$

where N denotes the total frames of the test signal.

Thus, the MPQoS can be used as a metric for ranking a homogeneous clip into a perceived-quality scale.

MEAN PQoS VS. BIT-RATE CURVES

The most significant encoding parameter with strong influence over the video quality, and the storage and network resource requirements is the encoding bit rate (i.e., the compression ratio), given that the frame rate and the picture resolution are not modified for a specific end-user terminal device. In order to identify the relation of the MPQoS with the encoding bit rate, four homogeneous and short-in-duration test sequences, which are representative of specific spatial and temporal activity levels, were used. Table 1 depicts these four video clips.

Each test video clip was transcoded from its original MPEG-2 format at 12 Mbps with PAL resolution at 25 fps to ISO MPEG-4 (simple profile) format at different constant bit rates (spanning a range from 50 Kbps to 1.5 Mbps). For each corresponding bit rate, a different ISO MPEG-4-compliant file with CIF resolution (352x288) at 25 fps was created.

Figure 1. The instant PQoS of the "Mobile & Calendar" clip (CIF resolution) derived by the QMS tool

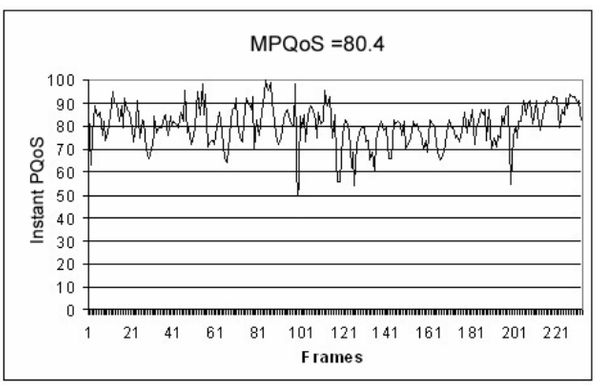
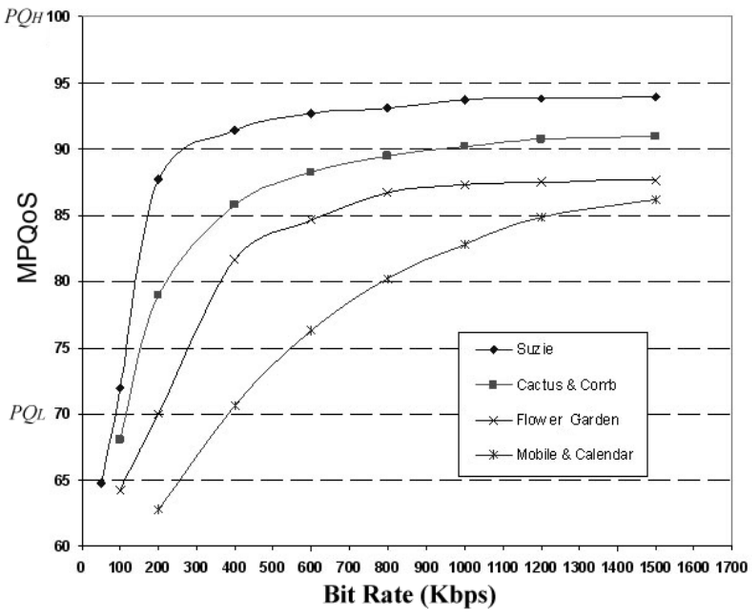


Table 1. Test video sequences

Clip 1	Low Spatial & Temporal Activity Level	Suzie	
Clip 2		Cactus	
Clip 3	High Spatial & Temporal Activity Level	Flower Garden	
Clip 4		Mobile & Calendar	

Figure 2. The MPQoS vs. bit rate curves for CIF resolution



Each ISO MPEG-4 video clip was then used as input in the QMS tool. From the resulting instant PQoS-vs.-time graph (like the one in Figure 1), the MPQoS value of each clip was calculated. This experimental procedure was repeated for each video clip in CIF resolution.

The results of these experiments are depicted in Figure 2, where PQ_L denotes the lowest acceptable MPQoS level (which is considered equal to 70 for this paper) and PQ_H denotes the best MPQoS level that each video can reach.

Comparing the experimental curves of Figure 2 to those resulting from the theoretical algebraic benefit functions described in Lee and Srivastava (2001) and Sabata, Chatterjee, and Sydir (1998), and to the contrast-response saturation curves (Wang et al., 2003), qualitative similarity among them is noticed. Thus, the experimental curves derived using the QMS tool are qualitatively very similar to what was theoretically expected, proving, therefore, their validity.

Moreover, it is of great importance the fact that the MPQoS vs. bit-rate curves is not identical for all the types of audiovisual content, but a differentiation among them lies on the S-T activity of the video content. Thus, the curve has low slope and transposes to the lower right area of the MPQoS-vs.-bit-rate plane for AV content of high S-T activity. On the contrary, the curve has high slope and transposes to the upper left area for low S-T activity content. In addition, when the encoding bit rate decreases below a threshold, which depends on the video content, the MPQoS practically collapses.

However, it should be noted that the MPQoS as a metric is valid for video clips that are homogeneous in respect to the S-T activity of their content, that is, for example, video clips whose contents are exclusively talk shows or football matches. For heterogeneous video clips, the method is not very accurate, producing MPQoS vs. bit-rate curves that are indistinguishable.

EXPONENTIAL APPROXIMATION OF MPQoS VS. BIT-RATE CURVES

The experimental MPQoS curves of Figure 2 can be successfully approximated by a group of exponential functions. Consequently, the MPQoS level of an MPEG-4 video clip, encoded at bit rate *BR*, can be analytically estimated by the following equation:

$$\text{MPQoS} = [PQ_H - PQ_L]\,(1 - e^{-\alpha[BR-BRL]}) + PQ_L, \quad \alpha > 0 \text{ and } BR > BR_L \qquad (4)$$

where the parameter α is the time constant of the exponential function, determines the shape of the curve.

Since the maximum deviation error between the experimental and the exponentially approximated MPQoS curves was measured to be less than 4% in the worst case (for all the test signals), the proposed exponential model of MPQoS vs. bit rate can be considered that approximates successfully the corresponding experimental curves.

Referring to this approximation, each MPQoS vs. bit-rate curve can be uniquely described by the following three elements:

1. The minimum bit rate (BR_L) that corresponds to the lowest acceptable PQoS level (PQ_L is considered equal to 70 for this paper)
2. The highest reached PQoS level (PQ_H)
3. A parameter α that defines the shape and, subsequently, the slope of the curve

So, a triplet (α, BR_L, PQ_H) can be defined that contains the QoS elements that are necessary for describing analytically the exponentially approximated MPQoS vs. bit-rate curve for a given video signal.

Experimental curves of MPQoS vs. bit rate and their corresponding exponential approximations were compared not only for the above four reference video clips, but also for other AV content, spanning a wide range of S-T activity. The results showed that the experimental curves of MPQoS vs. bit rate were successfully approximated by exponential functions with the triple elements being in the range of those in Table 2.

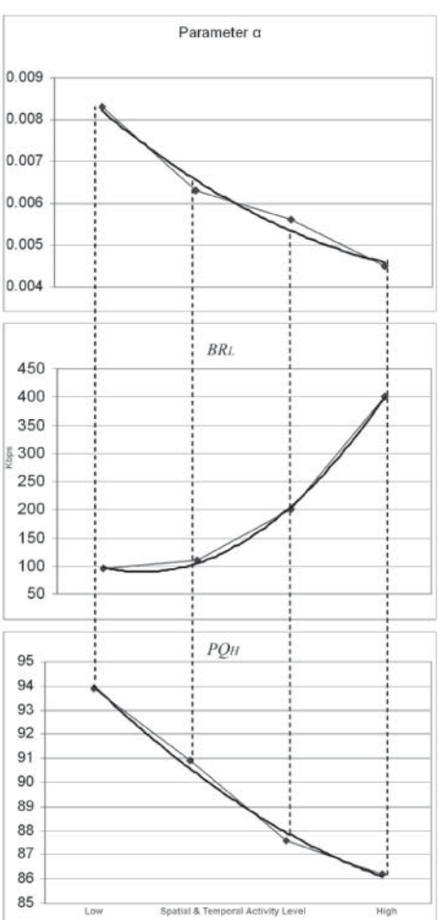
Table 2. Triple elements that correspond to the test signals (MPEG-4 / CIF)

Test Sequence	α	BR_L (Kbps)	PQ_H (Quality Units)
Suzie	0.0083	95	93.91
Cactus	0.0063	110	90.89
Flower	0.0056	200	87.62
Mobile	0.0045	400	86.20

FAST ESTIMATION OF THE TRIPLE ELEMENTS

The determination of the bit rates for a video service that correspond to the various quality levels can be achieved by the multiple repeating of postencoding measurements of the MPQoS at various bit rates. Since this is a complicated and time-consuming process, an alternative simple and fast preencoding

Figure 3. Variation of the triple elements (MPEG-4/CIF)



evaluation method is proposed based on the use of the triplet (α, BR_L, PQ_H).

The triple elements are not independent among them. On the contrary, there is a correlation that can be derived experimentally. Considering the four test video signals, the variation of their triple elements vs. the S-T activity level is depicted in Figure 3.

According to Figure 3, if one out of the three triple elements is specified for a given video clip, then the other two can be estimated graphically. Due to the exponential form of the MPQoS vs. bit-rate curves, the PQ_H element can be simply derived: Using the QMS tool, which was described in the background, only one measurement or estimation of the MPQoS at a high encoding bit rate is sufficient for the accurate determination of the PQ_H value for a given video clip.

Using the estimated PQ_H value as input in the reference curves of Figure 3, the corresponding values of BR_L and α can be graphically extrapolated by a vertical line that passes through the specific PQ_H value and cuts the other two curves at the corresponding BR_L and α values, which complete the specific triplet. Thus, having defined the complete triplet for a given clip, the analytical exponential expression of the MPQoS vs. bit rate can be deduced using Equation 4. This enables the preencoding of the MPQoS evaluation for a specific video signal because Equation 4 can indicate the accurate bit rate that corresponds to a specific PQoS level and vice versa.

FUTURE TRENDS

The ongoing offering of multimedia applications and the continuous distribution of mobile network terminals with multimedia playback capabilities, such as cellular phones, have resulted in the significant increment of network load and traffic. Due to this, the multimedia services that are delivered to an end-user terminal device often experience unexpected quality degradation resulting from network QoS-sensitive parameters (e.g., delay, jitter).

One of the current trends in the quality-evaluation research field is the correlation of these network parameters with perceived-quality degradation. This mapping will enable the evaluation of the network capability to deliver successfully a multimedia service at an acceptable PQoS level.

CONCLUSION

In this paper, the mean PQoS is proposed as a metric characterizing a homogeneous (in content) video clip as a single entity. The experimental MPQoS vs. bit-rate curves are successfully approximated by a group of exponential functions, with a deviation error of less than 4%, enabling the analytical description of the MPQoS curves by three elements. Based on this, a method for fast preencoding estimation of the MPQoS level is proposed that enables optimized utilization of the available storage and bandwidth resources because only the resources that are sufficient to maintain a specific level of user satisfaction are allocated.

The accuracy of the proposed assessment method depends on the homogeneity of the video content under consideration because only when the AV content is representative of a specific S-T activity level, like a talk show or a sports event, the corresponding MPQoS curves are well distinguished and successfully describe the PQoS characteristics of the clip.

This requirement for homogeneity limits the duration of the suitable video signals to low levels. However, this is not an obstacle for the upcoming 3G and 4G applications, where the duration of the multimedia services will be short.

ACKNOWLEDGEMENT

The work in this paper was carried out in the frame of the Information Society Technologies (IST) project ENTHRONE/FP6-507637.

REFERENCES

Alpert, T., & Contin, L. (1997). *DSCQE experiment for the evaluation of the MPEG-4 VM on error robustness functionality* (ISO/IEC – JTC1/SC29/WG11, MPEG 97/M1604). Geneva: International Organization of Standardization.

Bradley, A. P. (1999). A wavelet difference predictor. *IEEE Transactions on Image Processing, 5*, 717-730.

Daly, S. (1992). The visible difference predictor: An algorithm for the assessment of image fidelity. *Proceedings of Society of Optical Engineering, 1616*, (pp. 2-15).

Guawan, I. P., & Ghanbari, M. (2003). *Reduced-reference picture quality estimation by using local harmonic amplitude information.* London Communications Symposium 2003, London.

ITU-R. (1996). *Methodology for the subjective assessment of the quality of television pictures* (Recommendation BT.500-7, Rev. ed.). Geneva: International Telecommunication Union.

Lai, Y. K., & Kuo, J. (2000). A haar wavelet approach to compressed image quality measurement. *Journal of Visual Communication and Image Understanding, 11*, 81-84.

Lauterjung, J. (1998). Picture quality measurement. *Proceedings of the International Broadcasting Convention (IBC)*, 413-417.

Lee, W., & Srivastava, J. (2001). An algebraic QoS-based resource allocation model for competitive multimedia applications. *International Journal of Multimedia Tools and Applications, 13*, 197-212.

Lu, L., Wang, Z., Bovik, A. C., & Kouloheris, J. (2002). Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video. *IEEE International Conference on Multimedia*, Lausanne, Switzerland.

MPEG Test. (1999). *Report of the formal verification tests on MPEG-4 coding efficiency for low and medium bit rates* (Doc. ISO/MPEG N2826). Geneva: International Organization of Standardization.

Pereira, F., & Alpert, T. (1997). MPEG-4 video subjective test procedures and results. *IEEE Transactions on Circuits and Systems for Video Technology, 7*(1), 32-51.

Sabata, B., Chatterjee, S., & Sydir, J. (1998). Dynamic adaptation of video for transmission under resource constraints. *International Conference of Image Processing*, Chicago.

Tan, K. T., & Ghanbari, M. (2000). A multi-metric objective picture quality measurements model for MPEG video. *IEEE Transactions on Circuits and Systems for Video Technology, 10*(7), 1208-1213.

VQEG. (2000). *Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment*. Retrieved March 2000 from *http://www.vqeg.org*

Wang, Z., Bovik, A. C., & Lu, L. (2002). Why is image quality assessment so difficult. *Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing, 4*, 3313-3316.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 1-14.

Wang, Z., Lu, L., & Bovik, A. C. (2004). Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication, 19*(2), 121-132.

Wang, Z., Sheikh, H. R., & Bovik, A. C. (2003). Objective video quality assessment. In B. Furht & O. Marqure (Eds.), *The handbook of video databases: Design and applications* (pp. 1041-1078). CRC Press.

Watson, A. B., Hu, J., & McGowan, J. F. (2001). DVQ: A digital video quality metric based on human vision. *Journal of Electronic Imaging, 10*(1), 20-29.

Wolf, S., & Pinson, M. H. (1999). Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system. *SPIE International Symposium on Voice, Video, and Data Communications*, 11-22.

KEY TERMS

Benefit Function: Theoretical algebraic functions depicting the user satisfaction for a multimedia service in correlation with the allocated resources.

Bit Rate: A data rate expressed in bits per second. In video encoding, the bit rate can be constant, which means that it retains a specific value for the whole encoding process, or variable, which means that it fluctuates around a specific value according to the content of the video signal.

CIF (Common Intermediate Format): A typical video or image resolution value with dimensions 352x288 pixels.

Contrast Response Saturation Curves: Curves representing the saturation characteristics of neurons in the human visual system (HVS).

MPEG-4: Digital video-compression standard based on the encoding of audiovisual objects.

MPQoS (Mean Perceived Quality of Service): The averaged PQoS that corresponds to a multimedia service.

Objective Measurement of PQoS: A category of assessment methods that evaluate the PQoS level based on metrics that can be measured objectively.

PQoS (Perceived Quality of Service): The perceived quality level that a user experiences from a multimedia service.

Quality Degradation: The drop of the PQoS to a lower level.

Spatial-Temporal Activity Level: The dynamics of the video content in respect to its spatial and temporal characteristics.

The Online Discussion and Student Success in Web-Based Education

Erik Benrud
American University, USA

INTRODUCTION

This article examines the performance of students in a Web-based corporate finance course and how the technologies associated with communication on the Internet can enhance student learning. The article provides statistical evidence that documents that the online discussion board in a Web-based course can significantly enhance the learning process even in a quantitative course such as corporate finance. The results show that ex ante predictors of student performance that had been found useful in predicting student success in face-to-face classes also had significant predictive power for exam performance in the online course. However, these predictors did not have predictive power for participation in the online discussion. Yet, online participation and exam performance were highly correlated. This suggests that the use of the online discussion board technology by the students enhanced the performance of students who otherwise would not have performed as well without the discussion.

The online discussion in a Web-based course promotes active learning, and active learning improves student performance. Educators have long recognized the importance of an active learning environment; see Dewey (1938) and Lewin (1951). It is no surprise, therefore, that later research such as Dumant (1996) recognized the online discussion as one of the strengths of Web-based learning. Some researchers, such as Moore and Kearsley (1995) and Cecez-Kecmanovic and Webb (2000) have gone on to propose that the online discussion may even challenge the limits of the face-to-face (F2F) environment.

To explore the effect of the discussion on students' grades, we must first measure the amount of variation in the grades explained by ex ante measures that previous studies have used. The Graduate Management Aptitude Test[1] (GMAT) score, gender, and age were used. A variable that indicated whether the student considered himself or herself someone who took most courses on the Web, that is, a "Web student," was also included, and these four ex ante predictors of student performance explained over 35 percent of the variation of the final course grades in a sample of 53 students. This level of explanatory power using these predictors was similar to that of previous studies concerning F2F finance classes; see Simpson and Sumrall (1979) and Borde, Byrd, and Modani (1998). In this study, with the exception of the condition "Web student," these determinants were poor predictors of online discussion participation; however, there was a significant relationship between online discussion participation and performance on the exams. These results provide evidence that multimedia technologies that promote student interaction can aid the learning process in a course that is largely quantitative in nature.

THE ROLE OF THE ONLINE DISCUSSION

The Internet is ideally suited for a learning tool such as a discussion board where the students can interact and discover answers for themselves. The overall effect of this combination of computer and teaching technology appeared to stimulate student interest and enhanced the learning process. The data gathered in this study indicates that the students appreciated the use of the technology and that each student tended to benefit to a degree that was commensurate with his or her level of participation.

The online discussion consisted of a Socratic dialogue that was led by the instructor. This is an ancient technique that recognizes that student activity aids the learning process. As applied here, it is a learning technique that begins with a single question and then requires participants to continually answer a

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

series of questions that are generated from answers to previous questions with the goal of learning about a topic. The Socratic dialogue is widely used in F2F classes around the world, see Ross, (1993). Using the interactive technology of the discussion board over the Internet seemed especially beneficial. Having the discussion over a week's time on the Internet allowed students time to think and reflect both before and after their contribution. The students were motivated to participate because the discussion made up 25 percent of their final grade, which was equal to the weight of each of the two exams. The remaining 25 percent was earned from small assignments and one project.

The students earned a portion of the discussion grade each week. At the beginning of each week, a question would be posed such as: "Corporations must pay institutions like Moody's and S&P to have their debt rated. What is the advantage to the corporation of having its debt rated?" The students would post answers and, with the guidance of the instructor, would explore a number of related issues. The students earned credit by "adding value" to the dialogue each week. Students were invited to contribute reasoned guesses, personal anecdotes, and examples from the Internet. One well-thought-out and thorough contribution would earn a student a perfect score for the week. Several small contributions would earn a perfect score as well. The grades earned from discussion participation were generally good. The average discussion grade earned, as a percentage of total points, was 92.81 with a standard deviation of 8.75. The results were highly skewed in that nine of the 53 students earned 100 percent of the online discussion grade. The corresponding percent of total points earned for the course without the discussion had an average equal to 86.21 and a standard deviation equal to 7.26 for all students.

The students generally reacted favorably to the online discussion. All 53 students took a confidential survey that asked them questions about their perceptions of the online discussion. The results reveal that 60 percent felt that this course used the online discussion *more* than the average Web-course they had taken; 76 percent rated the quality of the discussion *higher* than the average they had experienced in other Web-classes; and 55 percent said that the online discussion *significantly aided* their understanding of corporate finance.

STATISTICAL ANALYSIS

To begin the analysis, this study used the variables gender, age, GMAT score, and whether a student was a Web-MBA student to explain performance in the course. Table 1 lists the correlations of various components of these ex ante characteristics with the grades and discussion-participation data. The variables are defined in the list below. The letter "N" appears at the end of a definition if the data for that variable has a bell-shaped or normal distribution, which means the test results for those variables are more reliable.[2]

- **AGE:** The age of the student at the beginning of the class; the range was 21 to 55 with a mean of 31.47, N.
- **DE:** Number of discussion entries, a simple count of the number of times a student made an entry of any kind in the discussion, N.
- **DISC:** Grade for student participation in the online discussion.
- **FAVG:** Final average grade for the course, N.
- **FINEX:** Final exam grade, N.
- **GEN:** Gender, this is a dummy variable where GEN=1 represents male and GEN=0 represents female, the mean was 0.540.
- **GMAT:** Graduate Management Aptitude Test score.
- **GWD:** Grade for the course without discussion, to get this the discussion grade was removed from the final average and that result was inflated to represent a score out of 100 percent, N.
- **MT:** Midterm exam grade, N.
- **PROJ:** Grade on a project that required the creation of a spreadsheet.
- **WC:** Word count; the total number of words the student wrote in the discussion over the entire course, the range was 391 to 5524 with a mean equal to 2164, N.
- **WMBA:** Whether the student considered him/herself a Web-MBA student as opposed to student who takes most courses in a F2F environment, WMBA=1 for Web-MBA students, else 0; the mean was 0.684.

For each pair of variables, Table 1 lists both the correlation coefficient and the probability value associated with a hypothesis that the correlation is zero. In those cases where an assumption of normality could not be rejected for both variables, the correlation and p-value on the table are in bold font. The correlation coefficient is a measure of the strength of the linear relationship between the variables.

Table 1 displays several interesting phenomena. AGE was positively correlated, but not at a significant level, with most measures of performance. The GMAT score served as a good predictor of test scores and the project score (symbols: MT, FINEX, and PROJ). The correlation of the GMAT score with the three measures of student participation in the online discussion was much weaker. Those three measures of student participation in the online discussion were the number of discussion entries, the word count, and the discussion grade for each student (symbols: DE, WC and DISC).

Some interesting observations concern the discrete binomial, or "zero/one variables," GEN and WMBA, and they are included on Table 1 for descriptive purposes. As found in previous studies, males had a higher level of success on exams. Students who consider themselves Web students, that is, WMBA=1, had a superior performance in all categories too.

Analysis of variance tests (ANOVA) allow us to determine if the effects of WMBA and GEN were statistically significant. Consistent with the requirements of ANOVA, Table 2 reports the results for the normally distributed measures of performance: FINEX, GWD, FAVG, DE, and WC. The condition

Table 1. Correlation matrix of grades, discussion data, and student characteristics

Correlation coef. with p-value underneath, e.g., corr(Disc,MT)=0.087 and p-value=0.460. Cells in BOLD indicate both variables pass tests for normality.											
	DISC	**MT**	**FINEX**	**PROJ**	**FAVG**	**GWD**	**DE**	**WC**	**GEN**	**AGE**	**WMBA**
MT	0.087										
	0.460										
FINEX	0.297	**0.673**									
	0.010	**0.000**									
PROJ	0.143	0.366	0.41								
	0.222	0.001	0.000								
FAVG	0.573	**0.755**	**0.88**	0.528							
	0.000	**0.000**	**0.000**	0.000							
GWD	0.281	**0.85**	**0.913**	0.564	**0.948**						
	0.015	**0.000**	**0.000**	0.000	**0.000**						
DE	0.513	**0.329**	**0.322**	0.104	**0.471**	**0.351**					
	0	**0.004**	**0.005**	0.374	**0**	**0.002**					
WC	0.515	**0.362**	**0.428**	0.19	**0.57**	**0.466**	**0.755**				
	0.000	**0.001**	**0.000**	0.102	**0.000**	**0.000**	**0.000**				
GEN	0.032	0.346	0.451	0.248	0.369	0.419	0.053	0.136			
	0.787	0.002	0.000	0.032	0.001	0.000	0.649	0.245			
AGE	0.099	**-0.013**	**0.093**	0.053	**0.124**	**0.107**	**0.13**	**0.169**	0.034		
	0.398	**0.914**	**0.427**	0.650	**0.288**	**0.362**	**0.265**	**0.147**	0.773		
WMBA	0.246	0.288	0.274	0.125	0.316	0.271	0.293	0.211	0.122	-0.052	
	0.034	0.012	0.017	0.285	0.006	0.019	0.011	0.069	0.298	0.657	
GMAT	0.178	**0.301**	**0.368**	0.404	**0.421**	**0.414**	**0.063**	**0.245**	0.509	**-0.171**	0.273
	0.203	**0.029**	**0.007**	0.003	**0.002**	**0.002**	**0.652**	**0.077**	0.000	**0.221**	0.048

Table 2. ANOVA results for dummy variables GEN and WMBA

F-statistic and probability value are reported in each cell.						
		FINEX	**GWD**	**FAVG**	**DE**	**WC**
GEN	F-stat.=	18.66	15.55	11.53	0.210	1.38
	p-value=	0.000	0.000	0.001	0.649	0.245
WMBA	F-stat.=	5.94	5.77	8.10	6.84	3.41
	p-value=	0.017	0.019	0.006	0.011	0.069

Table 3. Regression of student performance on ex ante variables

Results in each cell in the explanatory variables columns are the coefficient, (t-statistic), probability value. For example, for the first equation for FAVG, the intercept coefficient is 78.307, the t-statistic is 18.5, and the probability value is 0.000.								
Dependant Variable		**explanatory variables**					**R^2**	**F-stat**
		Constant	**GEN**	**AGE**	**WMBA**	**GMAT**	**adj.R^2**	**P-value**
FAVG	**coef=**	78.307	4.701		3.129	0.0105	0.347	8.690
	t-stat=	(18.5)	(2.92)		(1.81)	(1.20)	0.307	0.000
	p-val=	0.0000	0.005		0.077	0.234		
DISC		90.550	3.175			0.004	0.0740	2.000
		(24.0)	(1.58)			(0.57)	0.0370	0.146
		0.000	0.121			0.571		
GWD		74.412	5.163		3.968	0.013	0.331	8.096
		(14.5)	(2.72)		(1.86)	(1.21)	0.290	0.000
		0.000	0.009		0.069	0.231		
PROJ		86.983	1.670			0.018	0.183	5.600
		(21.2)	(1.05)			(2.34)	0.150	0.006
		0.000	0.300			0.023		
WC		-376.5		32.300	375.31	2.482	0.130	2.400
		(-0.29)		(1.41)	(1.42)	(1.39)	0.077	0.075
		0.774		0.165	0.161	0.169		

WMBA=1 had a positive effect in all categories, and the effect was significant at the 10 percent level in all five cases. The condition WMBA=1 was the one ex ante predictor that had a significant relationship with DE and WC, and it probably indicated those students who had more experience with Web-based activities. This points to how a student's familiarity with the learning technologies employed in a course will affect that student's performance.

The ANOVA results show that males had significantly higher scores for the final exam, the grades without the discussion grade, and the course grade (FINEX, GWD, FAVG). This is congruent with previous research. The reason for the lower level of significance of GEN with respect to FAVG is explained by the fact that there was not a significant difference in the student participation in the online discussion for males and females, and that discussion grade is included in FAVG. For the raw measures DE and WC, there was not a significant difference in the participation rates of males and females. We should also note that for the non-normally distributed variable DISC, the discussion grade, males only slightly outperformed females. The average grades for males and females were 93.1 and 92.5 respectively with an overall standard deviation of 8.75.

Ordinary least squares (OLS) regressions can measure predictive power of the ex ante variables. Table C lists the results of regressions of the final grades (FAVG), the discussion grades (DISC), the grades without the discussion (GWD), the project grade (PROJ), and the word count (WC) on the indicated variables.

The equations for FAVG and GWD had the highest explanatory power, and the equation for PROJ had significant explanatory power too. The explanatory power of the equations for WC and DISC

Table 4a. OLS regression of final exam on ex ante performance in the course

Results in each cell in the explanatory variables columns are the coefficient, (t-statistic), probability value.							
Dependant Variable	**explanatory variables**					**R^2 adj.R^2**	**F-stat**
	Constant	**GEN**	**MT**	**DISC**	**WC**		
FINEX	-21.70	6.010	0.766	0.334		0.564	30.653
	(-2.42)	(3.11)	(7.93)	(3.54)		0.546	P=0.000
	0.018	0.003	0.000	0.001			
FINEX	9.669	5.961	0.692		0.0026	0.545	28.386
	(1.05)	(3.01)	(5.84)		(2.48)	0.526	P=0.000
	0.300	0.004	0.000		0.0015		

Table 4b. Two-stage least squares estimation

Instrument list: C AGE AGE^2 AGE^{-1} WMBA GEN Results in each cell in the explanatory variables columns are the coefficient, (t-statistic), probability value.							
Dependant Variable	**explanatory variables**					**R^2 adj.R^2**	**F-stat. p-value**
	Constant	**Gen**	**GWD**	**WC**	**Age**		
WC	-8493		120.15		10.212	0.033	3.781
	(1.98)		(2.33)		(0.59)	0.006	0.027
	0.0518		0.023		0.555		
GWD	70.923	4.534		0.006		0.216	11.038
	(14.08)	(2.77)		(2.44)		0.194	0.000
	0.000	0.007		0.017			

were much lower; although the equation for WC was significant at the ten-percent level, the results for WC and DISC were not significant at the 5 percent level.

As we would expect from past research, GEN had very significant coefficients in the equations for FAVG and GWD, which means that the condition "male" was associated with higher final averages and grades without discussion. GMAT was only marginally significant in most cases, but this could be the result of the high correlation of GMAT with GEN.

We can use OLS regressions to demonstrate how online discussion performance, as measured by DISC and WC, affected FINEX because the discussion occurred before FINEX was determined. In a regression of FINEX on GEN, MT, and DISC, the coefficient for DISC had a t-statistic greater than that for GEN. These results are on Table 4a. Since WC was normally distributed and was a raw measure of effort, a second specification on Table 4a replaces DISC with WC. The t-statistic for the discussion variable decreased slightly, as did the coefficient of determination symbolized by R^2. Both t-statistics were significant, however, and both R^2 values exceeded 50 percent. The coefficient of determination is a measure of variation explained, which means that in this case the variables in each equation explained over half of the differences in the grades on the final exams.

Table 4b gives the result of a second set of equations, which used two-stage least squares (TSLS) to estimate the effect of WC on GWD and then GWD on WC. TSLS was required here because GWD and WC developed simultaneously during the course. The results show that each had a significant relationship with the other.

The purpose of this section has been to report the statistical results and point out the interesting relationships. Many of those relationships are congruent with earlier work. For example, male students and those who had higher GMAT scores had higher exam grades. The most interesting results concern the

grades for the online discussion, which had a low correlation with the ex ante student characteristics GEN and AGE but were highly correlated with exam grades. The next section discusses some of the implications of these results.

DISCUSSION OF EMPIRICAL RESULTS

Consistent with previous research concerning F2F finance classes, the following were significant ex ant predictors of performance: gender, GMAT score, and age. As we might expect for a Web-based course, students who considered themselves web students or Web-MBA students, performed significantly higher for most of the grade variables. The most interesting point is that the traditional ex ante characteristics did not predict performance in the online discussion very well, yet there was a strong relationship between the online discussion and exam grades.

Gender displayed a very weak relationship with measures pertaining to the online discussion. The word count was normally distributed and did not have a significant relationship with gender in either an ANOVA or OLS regression. Word count was an unrefined measure, but this was an advantage in that it served as a direct measure of effort, and it was unaffected by the subjective opinions of the instructor. Word count was significantly correlated with each of the student's class scores. In summary, gender, age, GMAT score, and whether a Web-MBA student explained success on exams. With the exception of whether a Web-MBA student, these predictors were not significantly correlated with performance in the online discussion. Although these variables had low explanatory power for the online discussion measures, there was a high correlation between performance in the online discussion and exam grades.

Using the discussion score or word count in an equation with the gender variable and the midterm exam grades explained over 50 percent of the variation of the final exam grades. In fact, the discussion grade's coefficient had a larger t-statistic than the gender variable. The coefficient for word count in the equation for the final exam grade has a significant coefficient equal to 0.0026. This means that for every 385 words of writing in the online discussion, on average, a student's final exam grade was about one point higher: 385*0.0026»1. The TSLS results on table D2 indicate the effect of word count on the grades without the discussion grade. The coefficient was significant and estimated to be 0.006. This means that for every 167 words, on average, there was an associated increase in the grade without the discussion of one point: 167*0.006=1.

The results of this study show that a student's use of a technology such as an Internet discussion board can enhance that student's performance in other areas of a course. The ex ante measures of gender, age, and GMAT were not useful in predicting who would participate in and thus benefit from the discussion board. Use of the discussion board technology by the students had a significant and positive effect on the grades earned on the exams. Furthermore, the fact that students who considered themselves Web-MBA students had superior performance means that training and experience in the use of multimedia technologies is important in order to allow students to benefit from such technologies to a greater degree.

REFERENCES

Borde, S., Byrd, A., & Modani, N. (1998). Determinants of student performance in introductory corporate finance courses. *Journal of Financial Education, Fall*, 23-30.

Cecez-Kecmanovic, D. & Webb, C. (2000). A critical inquiry into Web-mediated collaborative learning. In A. Aggarwal (Ed.), *Web-based learning and teaching technologies: Opportunities and challenges*. Hershey, PA: Idea Group Publishing.

Dewey, J. (1938). *Experience in education*. New York: Macmillan.

Dumant, R. (1996). Teaching and learning in cyberspace. *IEEE Transactions on Professional Communication, 39*, 192-204.

Lewin, K. (1951). *Field theory in social sciences*, New York: Harper and Row Publishers.

Moore, M. & Kearsley, G. (1995). *Distance education: A systems view*. Belmont, CA: Wadsworth Publishing.

Ross, G.M. (1993). The origins and development of socratic thinking. *Aspects of Education 49*, 9–22.

Simpson, W., Sumrall, G, & Sumrall, P. (1979). The determinants of objective test scores by finance students. *Journal of Financial Education*, 58-62.

KEY TERMS

Active Learning: Learning where students perform tasks, that is, post notes on a discussion board, to help in the learning process.

Coefficient of Determination: A statistical measure of how well predictive variables did indeed predict the variable of interest.

Correlation Coefficient: A statistical method of measuring the strength of a linear relationship between two variables.

Ex Ante Predictors of Student Performance: Student characteristics prior to beginning a class which have been determined to help forecast the relative performance of the students.

Online Discussion Board: Often called a "forum," it is a technology which allows students to interact by posting messages to one another at a particular URL on the Internet.

Probability Value: A statistical measure that attempts to assess whether an outcome is due to chance or whether it actually reflects a true difference; a value less than 5 percent means that a relationship very likely exists and the result probably did not occur by chance.

Regression: A statistical method of estimating the exact amount a variable of interest will change in reaction to another variable.

Socratic Dialogue: A learning technology which requires the participants to answer a series of questions to discover an answer or truth concerning a certain topic; the questions are typically generated spontaneously from previous answers.

Student Participation in the Online Discussion: The level to which a student contributed in an online discussion as measured, for example, by the number of words posted or the number of entries or the grade issued by an instructor.

Web Student: A student that plans to take the majority, if not all, of his or her classes in a particular program of study over the Internet.

ENDNOTES

1 The GMAT is a standardized entry exam that is administered on certain dates around the world. Most graduate business schools require students to take this test as part of the application process.

2 More specifically, the "N" signifies that we cannot reject a null hypothesis of the variable being distributed normally distributed based on a Kolmogorov-Smirnov test using a 5 percent level of significance. If "N" does not appear, then the null hypothesis of normality was rejected.

Open Source Intellectual Property Rights

Stewart T. Fleming
University of Otago, New Zealand

INTRODUCTION

The open source software movement exists as a loose collection of individuals, organizations, and philosophies roughly grouped under the intent of making software source code as widely available as possible (Raymond, 1998). While the movement as such can trace its roots back more than 30 years to the development of academic software, the Internet, the World Wide Web, and so forth, the popularization of the movement grew significantly from the mid-80s (Naughton, 2000).

The free software movement takes open source one step further, asserting that in addition to freedom of availability through publication, there should be legally-enforceable rights to ensure that it stays freely available and that such protections should extend to derived works (Stallman, 2002).

The impetus of both movements has resulted in the widespread distribution of a significant amount of free software, particularly GNU/Linux and Apache Web server. The nature of this software and the scale of installation appear to be an emerging concern for closed software vendors. At this time, we are seeing the emergence of legal challenges to the open source movement and a clash with the changing landscape of intellectual property and copyright protection.

There is spirited debate within and between both movements regarding the nature of open source software and the concerns over the extent to which software should remain free or become proprietary. This article concentrates on the issues directly relating to open source licenses, their impact on copyright and intellectual property rights, and the legal risks that may arise. For more general reference, the reader is directed to the Web sites of the Free Software Foundation (http://www.fsf.org), Open Source Initiative (http://www.opensource.org), and the excellent bibliography maintained by Stefan Koch (http://wwwai.wu-wien.ac.at/~koch/forschung/sw-eng/oss_list.html).

BACKGROUND

Motivations for Participation

The open source software movement is motivated by the desire to make software widely available in order to stimulate creative activity (either in the development of derivative software or in the use of that software in other endeavors). Free software requires open source software and goes further by pursuing the protection of ideas, ensuring that the intellectual basis for a software development can never be controlled exclusively or exploited.

Why would an individual decide to participate in a movement for which they might not accrue any direct financial benefit? Boyle (2003) discusses individual motivations with the notion of a reserve price, a level at which any individual decides to become an active participant rather than a consumer and to engage in some voluntary activity. It might be out of altruistic motives; it might be for the intellectual challenge; or it might be to solve a personal problem by making use of collaborative resources (and the entry level to collaboration is participation).

Gacek and Arief (2004) identify two additional motivations for open source participation: "developers are users" (p.35) and "knowledge shown through contributions increases the contributor's merit, which in turn leads to power" (p.37). This indicates a powerful motivation through self-interest and enhanced reputation with wide recognition of contributions, especially in large projects.

The presence of large industrial consortia in the open source movement and broad participation across many software development companies indicate that many commercial organizations also are motivated to participate. Table 1 lists broad categories to explain individual, academic, institution, and commercial motivations to participate in open source production activity.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Dempsey et al (2002) liken participation in open source software development to peer review in scientific research. By releasing one's software and using the software of others, continual innovations and improvements are made.

The large-scale collaborative nature of open source software development makes it important that the contributions of individuals are recognized, and the resulting situation is that ownership of any piece of open source software is jointly held. The solution that has evolved in the development of free software and open source movements has been the development of a variety of licensing models to ensure recognition and availability of contributions.

Open Source Definition

The Open Source Initiative (OSI) was begun in 1998 to make the case for open source software development to be more accessible to the commercial world. It provides samples of open source licenses and ratifies many of the licenses that cover various open source software developments. The Open Source Definition (Perens, 1999) is a useful description of the characteristics of what constitutes open source software (Table 2).

Software licenses that meet this definition can be considered as open source licenses, and the OSI provides certification for conforming licenses. While there have been many open source and free software licenses that have been created to suit various purposes, there are three main influences that will be considered in this article: GNU General Public License models, BSD license models, and Mozilla Public License (MPL) models.

MAIN FOCUS: OPEN SOURCE LICENSE MODELS

One of the most important developments to come out of the open source movement has been the proliferation and deep consideration of various licensing models to grant various rights to users of software. The collaborative depth of the movement is neatly illustrated by the spirited debate that surrounds issues that affect the community as a whole, and the diversity of the community provides broad viewpoints that cover all aspects, from the deeply technical to the legal. There is a wide range of different licenses (http://www.fsf.org/licenses/license-list.html), some free software, some not-free, and others incompatible with the General Public License (GPL). Table 2 summarizes the restrictions on various development activities applied by the three common classes of license.

BSD

The modified Berkeley Systems Development (BSD) license is an open source license with few restrictions and no impact on derived work. It requires only that attribution of copyright be made in source code and binary distribution of software. It specifically excludes any software warranties and disallows the use of the original organization in any advertising or promotion of derived works.

MIT(X11)

MIT(X11) is another open source license with very few restrictions and no impact on derived works. It requires only that a copyright notice be included with copies or substantial extracts of the software and excludes warranties.

The risk with unrestricted licenses such as BSD and MIT models is that a licensee can produce a derived work and not release improvements or enhancements, which might be useful to the wider community (Behlendorf, 1999).

Mozilla Public License

The modified version of the Mozilla Public License (MPL) (http://www.opensource.org/licenses/mozilla1.1.php) is a free software license that meets the OSI definition and is compatible with the GPL. It contains a number of complex provisions, but the inclusion of a multiple licensing clause allows it to be considered compatible with the GPL. The license is the controlling license for the Netscape Mozilla Web browser and associated software. It was developed specifically for the business situation at Netscape at the time of release but has since been used in many open source developments. The MPL/GPL/LGPL tri-license (http://www.mozilla.org/MPL/boilerplate-1.1/mpl-tri-license-txt) provides the mechanism for maintaining compatibility with the GPL.

The license includes clauses that are intended to deal with the software patent issue where source code that infringes on a software patent is deliberately or inadvertently introduced into a project. Behlendorf (1999) points out that there is a flaw in the waiver of patent rights in the license but suggests that, in general, the license is strong enough to support end-user development.

GNU General Public License

The GNU General Public Licenses (GNU GPL or GPL, for short) was originally developed by Richard Stallman around 1985 with the specific intention of protecting the ideas underlying the development of a particular piece of software. Free software does not mean that software must be made available without charge; it means that software, once released, must be always freely available. The GPL is a free software license that incorporates the "copyleft" provision that makes this freedom possible.

"To copyleft a program, we first state that it is copyrighted; then we add distribution terms, which are a legal instrument that gives everyone the rights to use, modify, and redistribute the program's code or any program derived from it but only if the distribution terms are unchanged. Thus, the code and the freedoms become legally inseparable" (http://www.gnu.org/copyleft/copyleft.html).

The GPL has provoked much debate, and the deliberate inclusion of political overtones in the wording of the license makes it unpalatable to some. Indeed, the Lesser GNU Public License (LGPL) is essentially the same as the GPL, but without the copyleft provision. This makes a free software license option available to commercial software developers without the obligation to release all of their source code in derived works.

In March 2003, the SCO Group, based in Utah, initiated a lawsuit against IBM alleging that proprietary SCO Linux code had been integrated into Linux, the leading open source operating system, and seeking damages since IBM has non-disclosure agreements in place with SCO regarding UNIX source code. The SCO Group also has sent letters to more than 1,500 large companies, advising them that they may face legal liability as Linux customers under the terms of the GPL. It is of great interest that the SCO-IBM lawsuit specifically targeted the GPL, which links source code with a legally protected freedom to distribute and make use of in derived work.

Whatever the motivations behind the lawsuit and its eventual outcome, as part of risk management activity, developers should be aware of the implications of creating and using software that is covered by the various licenses (Välimäki, 2004).

Adopting licenses other than the GPL weakens both it and the overall argument in favor of free software. This, in fact, may be the intention of the lawsuit—to mount a legal challenge that, if successful, would strongly dissuade developers from using the GPL.

Another barrier to the proliferation of open source software lies in the need to create broad-based standards. Without standards, interoperation of software created by multiple developers is difficult to achieve. However, the presence of patents that protect particular software inventions raises problems for the adoption of standards by the open source community. If there is only one way to accomplish a certain outcome, and if that method is protected, then development of an open source version is effectively blocked. Even if patents are licensed by their owners specifically for use in the development of open standards, there is an incompatibility with the GPL regarding freedom to create derived works (Rosen, 2004).

CRITICAL ISSUES AND FUTURE TRENDS

There are several issues that emerge from the consideration of open source and free software. As open source and free software becomes more widely used in different situations, potential legal risks become greater. Software development organizations must be aware of the possible effects of open source licenses when they undertake open source development.

The wide participation required by large-scale open source software development raises the risk of infringement on intellectual property, copyright, or software patent. The exclusion of warranties for software defects in most open source software licenses should cause organizations considering the adoption of open source software to carefully consider how quality and reliability can be assured.

The World Intellectual Property Organization survey of intellectual property on the Internet (WIPO, 2002) identifies open source software as the source of emergent copyright issues. It does not give any special treatment to the moral rights of authors with respect to software, and such rights are variable across international jurisdictions (Järvinen, 2002). Since the enhancement of reputation is an important motivating factor in participation in open source software development, software authors might benefit from more uniform international recognition of their right to assert authorship and their right to avoid derogatory treatment as author of a work.

Quality and reliability characteristics of open source software raise concerns for organizations in areas where certification is needed, such as in mission-critical activities like medicine. Harris (2004) provides an interesting account of how open source software was incorporated into the mission-critical data analysis tools for the Mars rovers *Spirit* and *Opportunity*. Zhao and Elbaum (2003) report that, although there was wide user participation in open source software projects that they surveyed, and although tools to track software issues were commonly used, the nature of testing activities was often shallow and imprecise. The lack of formal tools for testing, especially test coverage and regression testing, should lend a note of caution to those considering the use of open source software. The onus is on software developers making use of open source software to be duly diligent in their testing and integration of software.

A significant potential risk to open source software development is the protection of closed software markets by enforcement of software patents. An organization that has been granted a software patent for some algorithm or implementation is granted the rights to charge royalties for use, or it may force others to cease distribution of software that employs the scheme covered by the patent. Open source software is vulnerable to this form of restriction since all source code is publicly available. On the other hand, the distributed nature of the open source community can be a buffer against this form of restriction (Järvinen, 2002).

If we consider free software, the terms of the GPL have been written with reference to US law. Work is required to validate the terms of the license with respect to other jurisdictions. The main concern with the GPL is the copyleft clause covering derivative works. Järvinen (2002) has considered the GPL with respect to Finnish law; Välimäki (2001) gives a good account of the differences between US and European Union treatment of derivative works. Metzger and Jaeger (2001) have found that, although the GPL is generally compatible with German law, there may be issues with the complete exclusion of warranties. This may be the case in other jurisdictions where consumer protection laws are in force (e.g., US, EU, Finland, New Zealand) and warranties cannot be excluded. In the US, the lawsuit SCO vs. IBM in March 2003 is seen by many as a direct challenge to the GPL.

The exact nature of derivative works is determined by the courts. Välimäki (2004) summarizes different interpretations for what constitutes a derivative work (Table 4). Many of the issues regarding what does or does not constitute a derivative work are held only by mutual agreement among those in the open source software community. Software development organizations must be aware of the implications of open source software licenses, not only to cover the software that they distribute, but also those that cover any software they might use in the development. There is a serious risk of inadvertent breach of the GPL where an organization uses software covered by the GPL in proprietary software that it develops. Until there is a firm legal resolution in favor of or against the terms of the GPL, there is no firm basis for the application of the principles underlying the GPL.

In more general terms, the exact nature of security and liability with regard to open software is hard to establish. Kamp (2004) provides an interesting anecdote about the unimagined scale of distribution of a single piece of open source software. Although one of the much-vaunted strengths of the open source community is that "many eyes make all bugs shallow" (Raymond, 2001, p. 30), security issues still may be difficult to identify and resolve (Payne, 2002). Peer-review of public software is an advantage, but successful outcomes still depend on the motivation of properly skilled individuals to methodically study, probe, and fix open source software problems.

CONCLUSION

The future for open source licenses will be determined by the outcomes of legal challenges mounted in the coming years. The interpretation of many aspects of

the GPL only can be clarified properly through the courts of law. The interpretation in various jurisdictions will affect the international applicability of such licenses. Such tests are to be welcomed—they either confirm the strength of the open source and free software movements or, through a competitive influence, they cause them to reorganize in order to become stronger.

REFERENCES

Behlendorf, B. (1999). Open source as a business strategy. In C. DiBona, S. Ockman, & M. Stone (Eds.), *Open sources: Voices from the open source revolution*. Sebastopol, CA: O'Reilly & Associates.

Boyle, J. (2003). The second enclosure movement and the construction of the public domain. *Law and Contemporary Problems, 66*(33), 33-74.

Dempsey, B.J., Weiss, D., Jones, P., & Greenberg, J. (2002). Who is an open source software developer? *Communications of the ACM, 45*(2), 67-72.

Gacek, C., & Arief, B. (2004). The many meanings of open source. *IEEE Software, 21*(1), 34-40.

Harris, J.S. (2004). Mission-critical development with open source software: Lessons learned. *IEEE Software, 21*(1), 42-49.

Järvinen, H. (2002). Legal aspects of open source licensing. Helsinki, Finland: University of Helsinki, Department of Computer Science.

Kamp, P.-H. (2004). Keep in touch! *IEEE Software, 21*(1), 45-46.

Metzger, A., & Jaeger, T. (2001). Open source software and German copyright law. *International Review of Industrial Property and Copyright Law, 32*(1), 52-74.

Naughton, J. (2000). *A brief history of the future.* London: Phoenix Press.

Payne, C. (2002). On the security of open source software. *Information Systems Journal, 12*(1), 61-78.

Perens, B. (1999). The open source definition. In C. DiBona, S. Ockman, & M. Stone (Eds.), *Open sources: Voices from the open source revolution*. Sebastopol, CA: O'Reilly & Associates.

Ravicher, D. (2002). *Software derivative work: A circuit dependent determination*. New York: Patterson, Belknap, Webb and Tyler.

Raymond, E.S. (2001). *The cathedral and the bazaar: Musings on Linux and open source by an accidental revolutionary*. Sebastopol, CA: O'Reilly and Associates.

Rosen, L. (2004). *Open source licensing, software freedom and intellectual property law*. New York: Prentice Hall.

Stallman, R. (2002). *Free software, free society: Selected essays of Richard M Stallman*. GNU Press.

Välimäki, M. (2001). GNU general public license and the distribution of derivative works. *Proceedings of the Chaos Communication Congress*, Berlin, Germany.

Välimäki:, M. (in press). *A practical approach to the problem of open source and software patents*. European Intellectual Property Review.

Webbink, M.H. (2004). Open source software. *Proceedings of the 19th Annual Intellectual Property Conference*, Washington, D.C.

World Intellectual Property Organization. (2002). Intellectual property on the Internet: A survey of issues (No. WIPO/INT/02). Geneva, Switzerland.

Zhao, L., & Elbaum, S. (2003). Quality assurance under the open source development model. *Journal of Systems and Software, 66*, 65-75.

KEY TERMS

Assertion of Copyright: Retention of the protection right of copyright by an individual and, hence, the ability to collect any royalties that may be apportioned.

Attribution: Source code published under this license may be used freely, provided that the original author is attributed.

Copyleft: Provision in the GNU General Public License that forces any derived work based on software covered by the GPL to be covered by the GPL; that is, the author of a derived work must make all

source code available and comply with the terms of the GPL.

Copyright: Protected right in many jurisdictions that controls ownership over any material of a creative nature originated by an individual or organization.

Freely Available: Wide distribution at no cost to consumer.

Free Software: Software that is distributed under the terms of a license agreement that makes it freely available in source code form. Strong advocates of free software insist that the ideas underlying a piece of software, once published, must always be freely available.

General Public License: Specifically links source code to legally protected freedom to publish, distribute, and make use of derived works.

Intellectual Property (IP): Wider right to control ownership over any material of a conceptual nature (i.e., invention, idea, concept) as well as encompassing material originally covered by copyright.

Licensing Domain: Characterization of the breadth of availability and level of access to open materials.

Open Source: (See open source licensing model for strategies).

Open Source Licensing Model: A statement of the rights granted by the owner of some piece of open source software to the user.

Open Source Software: Computer software distributed under some license that permits the user to use, modify (including the creation of derived works), and distribute the software and any derived work, free from royalties.

Ownership: The association of the rights over intellectual property either with an institution or an individual so as to enable exploitation of that IP.

Ownership by Contract: Transfer or otherwise licensing all or part of copyright from the owner to one or more other parties covered by an explicit contract.

Public Domain: Owned by the public at large.

Publicly Available: Obtainable for free or minimum cost of materials.

Shareware: Software is available to users only on payment of a nominal fee.

Standard Setting: Source code under such a license may be used only in activity that defines an industry standard.

Work for Hire: An individual employed by an institution produces materials that are owned by the institution.

Open Source Software and International Outsourcing

Kirk St.Amant
Texas Tech University, USA

Brian Still
Texas Tech University, USA

INTRODUCTION

The popularity of open source software (OSS) has exploded among consumers and software developers. For example, today, the most popular Web server on the Internet is Apache, an open source product. Additionally, Linux (often considered one of the perfect examples of OSS) is now contesting Microsoft's dominance over the operating system market. OSS' flexibility, moreover, has allowed it to become a key international technology that could affect developments in global business practices. Despite these beneficial aspects, there are those who would claim it is difficult to implement and its core developers are undependable hobbyists. The purpose of this paper is to provide the reader with an overview of what OSS is, to present some of the benefits and limitations of using OSS, and to examine how international growth in OSS use could affect future business practices. By understanding these factors, readers will gain a better understanding of it and how OSS can be integrated into their organizational computing activities.

BACKGROUND

The driving force behind OSS is the Open Source movement, which can best be understood by what it opposes (proprietary software) and also what it supports (open software development). OSS advocates believe in an open exchange of ideas, an open coordination if not merging of different software, and, at the most crucial and basic level, an open access to the source code of software. In fact, Open Source creator Bruce Perens refers to it as a "bill of rights for the computer user" (Perens, 1999, p.171). Perens helped found the Open Source Initiative (OSI) in 1999, and only those software licenses that adhere to the guidelines of the OSI Open Source definition can use the trademark. OSI also maintains the Open Source definition and its registered trademark, and it campaigns actively for the Open Source movement and strict adherence to its definition.

The entire OSI Open Source definition can be viewed online at http://www.opensource.org/docs/definition.php. Its key tenets, however, can be summarized here: for a software license to be considered Open Source, users must have the right to make and even give away copies of the software for free. Additionally, and perhaps most importantly, OSS users must have the right both to view and to repair or modify the source code of the software they are using (Perens, 1999).

To appreciate the benefits and the limitations of OSS, one also must understand how it differs from proprietary software. In essence, the distinction has to do with differences in source code—the computer programming that tells software how to perform different activities or tasks. The motivation for this difference has to do with profits.

Proprietary software companies close access to the source code of their applications, because they consider it intellectual property critical to their business infrastructure. That is, once the programming of a software product is complete, these companies perform one final step, which is to prevent users from being able to see or to access the actual computer coding/programming that allows the software to operate. If any user could change the source code of the software, there eventually could be many different versions of it not easily supported by computers. If the user who purchased the software could change the source code, the user would not

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

need to pay the software company to make the change. With unrestricted access to the source code, a user even could develop another version of the software and then distribute it either at a lower cost or for free (Nadan, 2002).

According to the OSS model, the profitability of the software itself is not important. This is not to say that some OSS companies do not make money, for many do profit from providing services or support to users. The RedHat company (http://www.redhat.com), for example, makes a decent profit packaging and distributing Linux to users. While any user can download and install Linux for free, RedHat has convinced many users that by paying a fee to RedHat, they will get a guaranteed, ready-to-go version of Linux that comes with experienced support, such as training, manuals, or customer service (Young, 1999).

Additionally, OSS source code is not the intellectual property of one company or one programmer. Rather, it is more like community property that belongs to every user. With barriers removed as to who can access it and who cannot, the thinking behind this key Open Source tenet is that the more individuals who look at and modify the source code, the better that code will become. More bugs will be caught, more enhancements will be added, and the product will improve more quickly, as the experience and talents of a large community of developers is put to work making it better (Raymond, 1999).

This approach to software development and distribution has successfully threatened proprietary software's hold over the market in recent years. Although this OSS model seems revolutionary, it is actually the way things are done, according to Alan Cox, "in almost all serious grown up industry" (Cox, 2003, para. 11). In every field, consumers can go elsewhere if vendors are not supportive. In the auto industry, for example, individuals can choose the car they want from the dealer they want; they can look for the best deal, and they can even save money by fixing the car themselves (Cox, 2003). Because of OSS, software consumers now have that same sort of power. Instead of just one choice, one kind of license, and one price, consumers now have a choice of brand names, a chance to test multiple products for the right fit and buy, and, ultimately, the right to tinker with the software's source code on their own to make it work for their needs (Cox, 2003).

Just as Open Source wants to contribute to the public good, it also wants to put a flexible, more practical face on free software. Faced with losing the war for the hearts and minds of software users, the Open Source movement sacrifices the religious zeal of copyleft (preventing makers or modifiers of OSS from claiming ownership of and control over that programming) for a software certifying system that enables more software companies to license their work as Open Source (i.e., leaving the source code of their applications available and modifiable). In other words, OSI does not see itself in an antagonistic relationship with the software industry. Rather, "commercial software… [is] an ally to help spread the use of Open Source licensing" (Nadan, 2002, "The free/open source movement," para. 5).

To facilitate this relationship, OSI argues that business has much to gain from OSS. Business can, for example, outsource work to OSS developers and thus save money on in-house development. Additionally, a small business quickly can become the next Linux by interesting OSS developers in a project it has begun (Nadan, 2002). Almost overnight, scores of developers around the world could be working for free to make the project a reality.

Open Source, therefore, is about the true believers in free software trying to convince individuals in business to be believers, too. Why do they want business to use OSS? Because innovation, research, and development of software, once found primarily at big universities, is now carried out primarily in business. If business adopts OSS, its popularity not only will increase, but its quality will improve as more dollars and developers become dedicated to improving it. The question then becomes, How does one know if OSS is the right choice for his or her organization? To make an informed decision related to OSS, one needs to understand the benefits and the limitations of such programming.

MAIN FOCUS OF THE ARTICLE

Despite the inroads OSS has made in operating systems and Web servers, many businesspersons are still standoffish toward it. Others, having heard positive and negatives stories about OSS, are curious about what it can really do in comparison to proprietary software. By examining the strengths

and weaknesses of OSS and comparing it to proprietary software, one can establish the knowledge base needed to determine the specific situations where implementing OSS is the right decision.

OSS Strengths

- **Free Access to Source Code:** Organizations, especially those with skilled developers, can take advantage of free access to source code. OSS code is always available for modification, enabling developers to tinker with it to make it better for all users or just to meet their own needs or those of their organizations.
- **Costs:** A number of countries struggling economically, such as Taiwan and Brazil, have adopted OSS to save money (Liu, 2003). Many businesses faced with decreasing IT operating budgets but increasing software maintenance and licensing fees also have made the move toward OSS. Although there are indirect costs incurred using OSS, such as staff salary and training, proprietary software has these same costs. The fact that OSS starts free is a big plus in its favor.
- **Rapid Release Rate:** In the proprietary software model, software is never released until it is ready. If changes need to be made after the product is released, these alterations are not made and deployed as soon as possible. Rather, they are held back until the related company can be sure that all the bugs are fixed. OSS, however, works differently. As Raymond (1999) points out, updates to OSS are "released early and often," taking advantage of the large developer community working on the OSS to test, debug, and develop enhancements (p. 39). All of this is done at the same time, and releases are sometimes done daily, not every six months, so the work is efficient, and the improvement to the software is rapid.
- **Flexibility:** Open access to the source code gives users flexibility because they can modify the software to meet their needs. The existence of OSS also gives users flexibility simply because they have a choice that might not have existed before. They do not have to use proprietary software if there is OSS that works just as well. OSS provides further flexibility for those users that need to move to a new system. Rather than being stuck with "nonportable code and…forced to deal with whatever bugs" that come along with the software, they can use OSS that is "openly specified… [and] interchangeable" (Brase, 2003, vendor flexibility, para. 1).
- **Reliability:** Because OSS is peer-reviewed, and modifications are released quickly, problems with the software are caught and corrected at a rate countless times faster than that offered by proprietary software. It is not an industry secret that Linux, for example, is much more reliable than Windows. Exposed to the prying eyes of literally thousands of developers, Linux and other OSS are constantly being tested and tweaked to be made more crash proof. This tweaking also extends the life of the software, something which cannot be done with proprietary software unless users are willing to pay for upgrades. Many organizations have found themselves sitting on dated software and facing an expensive relicensing fee to get the new version. OSS can be refitted by the organization, if it is not already tweaked by the community of developers working on it. The software could be abandoned, and this has occurred before with OSS. But it has happened as well with proprietary software.
- **Developer Community:** In the end, the developer community is the greatest strength of OSS and one that proprietary software companies cannot match. Not all developers working on any given OSS project actually write code. But literally thousands upon thousands working on larger projects test, debug, and provide constant feedback to maintain the quality of the OSS. They are not forced to do it, but they contribute because they are stimulated by the challenge and empowered by the opportunity to help build and improve software that provides users, including themselves, with a high quality alternative to proprietary software.

OSS Weaknesses

Critics of OSS point out a number of deficiencies that make OSS too risky of a proposition to use in any sort of serious enterprise.

- **Loosely Organized Community of Hhobbyists:** It is a very real possibility that an OSS project could lose its support base of developers, should they get bored and move on to other projects. Although many that work on OSS are paid programmers and IT professionals, they often work on OSS outside of normal business hours. Many in business feel that professional developers working for companies that care about the bottom line in a competitive software market always will produce better software. They will stick around to support it, and the company will put its name on the line and stand behind it. In truth, the numbers of developers getting paid to work on OSS is increasing, and nearly one-third is paid for their work. In 2001, IBM, now a strong supporter of OSS, had around 1,500 of its developers working on just one OSS application—Linux (Goth, 2001).
- **Forking Source Code:** Source code is said to fork when another group of developers creates a derivative version of the source code that is separate, if not incompatible, with the current road the source code's development is following. The result is source code that takes a different fork in the road. Because anyone can access and modify OSS source code, forking has always been a danger that has been realized on occasions. The wide variety of operating systems that now exists, based on the BSD operating system, such as FreeBSD, OpenBSD, and NetBSD, serves as one example (DiBona, Ockman & Stone, 1999). Raymond (1999) argues that it is a taboo of the Open Source culture to fork projects, and in only special circumstances does it happen. Linux has not really forked, despite so many developers working on it. Carolyn Kenwood (2001) attributes this to its "accepted leadership structure, open membership and long-term contribution potential" (p. xiv). The GPL license, which Linux uses, is also a major deterrent to forking because there is no financial incentive to break off, since the forked code would have to be freely available under the terms of the license. Overall, however, forking is a legitimate potential weakness for OSS.
- **Lack of Technical Support:** In *CIO* magazine's 2002 survey of IT executives, "52 percent said a lack of vendor support was open source's primary weakness" (Koch, 2003, p. 55). Very rarely is software ever installed without some kind of hitch. In smaller organizations, the staff's depth of knowledge may not go deep enough to insure that support for the software can be taken care of internally. Because so many of the systems and applications that organizations run these days operate in hybrid environments where different tools run together on different platforms, technical support is crucial. Proprietary companies argue that Open Source cannot provide the technical support business expects and needs. There is no central help desk, no 1-800 number, no gold or silver levels of support that organizations can rely on for assistance. Recognizing that OSS must mirror at least the traditional technical support structure of proprietary models to address this perceived weakness, a number of "major vendors such as Dell, HP, IBM, Oracle and Sun" are beginning to support OSS (Koch, 2003, p. 55).
- **Lack of Suitable Business Applications:** Literally hundreds, if not thousands, of OSS applications can be downloaded for free off the Internet from sites like the Open Source Directory (http://www.osdir.com) or SourceForge (http://www.sourceforge.net). But a fair knock against OSS in the business world is that, aside from Linux and a few others, most OSS lacks the quality, maturity, or popularity to make business want to switch from the proprietary products it currently uses. Some think this is because building a word processor just is not sexy enough for OSS developers (Moody, 1998). While it may be changing, the nature of OSS is that those projects that developers choose to participate in are the ones that interest them, not necessarily those that others want done. If more companies begin to pay their developers to work on OSS, this situation may change. For now, however, OSS lacks the killer app for the desktop that matches Linux's impact on operating systems or Apache's on Web servers. OpenOffice, mentioned earlier, is an OSS alternative to Windows Office, but its user interface lacks

the sophistication and ease-of-use of Office, and so business has been slow to warm up to it. Until it or another OSS desktop application comes along that can seriously challenge Windows' lock on the desktop, those in decision-making positions will still not see OSS "as a legitimate alternative to proprietary software" (Goth, 2001, p. 105).

FUTURE TRENDS

While OSS might seem a bit of a novelty at this point in history, that perception is poised to change and to change rapidly in response to international business opportunities. The international use of proprietary software has long been plagued by two factors: cost and copyright. From a cost perspective, the for-profit nature of proprietary software has made it unavailable to large segments of the world's population. This situation is particularly the case in developing nations where high purchase prices often are associated with such materials. This cost factor not only restricts the number of prospective overseas consumers, but it also limits the scope of the viable international labor pool companies can tap.

The situation works as follows: Many developing nations possess a highly-skilled, well-trained workforce whose members can perform a variety of specialized technical tasks for a fraction of what it would cost in an industrialized nation. The savings that companies could incur through the use of such workers has prompted many organizations to adopt the practice of international outsourcing, a process in which technical work is sent to workers in developing nations. As a result, many companies have adopted such practices, and the degree of knowledge work that is expected to be outsourced in the near future is impressive.

For such outsourcing practices to work, employees in developing nations must have access to:

- the technologies needed to communicate with the overseas client offering such work; and
- the tools required to perform essential production tasks or services.

In both cases, this means software. That is, since much of this outsourcing work is conducted via online media, workers in developing nations need to have certain online communication software if they are to actually do work for clients in industrialized nations. Additionally, much of this work requires employees to use different digital tools (i.e., software packages) to perform essential tasks or develop desired products. The costs associated with proprietary software, however, mean that organizations often cannot take advantage of the full potential of this overseas labor force, for these prospective employees simply cannot afford the tools needed to engage in such activities.

This limited access to software also affects the consumer base that organizations can tap into in developing nations. In instances where the requisite software is available (e.g., India and China), outsourcing work (and outside money) moves into these counties and contributes to an economic boom. As a result, the middle classes of these nations begin to expand, and the members of this class increasingly have the financial power to purchase products sold by international technology companies. China, for example, has emerged as one off the world's largest markets for cellular telephones with some 42 million new mobile phone accounts opening in the year 2000 (China's economic power, 2001). Moreover, China's import of high-tech goods from the U.S. has risen from $970 million USD in 1992 to almost $4.6 billion USD in 2000 (Clifford & Roberts, 2001). Similarly, outsourcing has allowed India's growing middle class to amass aggregate purchasing power of some $420 billion USD (Malik, 2004). Thus, the more outsourcing work that can move into a developing nation, the more likely and the more rapidly that nation can become a market for various products.

One way to take advantage of this situation would be for companies to provide prospective international workers with access to free or inexpensive software products that would allow them to participate in outsourcing activities. Such an approach, however, would contribute to the second major software problem—copyright violation. In many developing nations, copyright laws are often weak (if not non-existent), or governments show little interest in enforcing them. As a result, many developing nations have developed black market businesses that sell pirated versions of software and other electronic goods for very low prices. Such

piracy reduces consumer desire to purchase legitimate and more costly versions of the same product, and thus affects a company's profit margins within that nation. Further complicating this problem is the fact that it is often difficult for companies to track down who is or was producing pirated versions of their products. Thus, while the distribution of cheap or free digital materials can help contribute to outsourcing activities, that same strategy can undermine an organization's ability to sell its products abroad.

Open source software, however, can offer a solution to this situation. Since it is free to use, OSS can provide individuals in developing nations with affordable materials that allow them to work within outsourcing relationships. Moreover, the flexibility allowed by OSS means that outsourcing workers could modify the software they use to perform a wide variety of tasks and reduce the need for buying different programs in order to work on different projects. As the software itself is produced by the outsourcing employee and not the client, concerns related to copyright and proprietary materials no longer need to be stumbling blocks to outsourcing relationships. Thus, it is perhaps no surprise that the use of OSS is growing rapidly in many of the world's developing nations (Open Source's Local Heroes, 2003).

This increased use of Open Source could contribute to two key developments related to overseas markets. First, it could increase the number of individuals able to work in outsourcing relationships, and thus increase the purchasing power of the related nation. Second, OSS could lead to the development of media that would allow poorer individuals to access the Internet or the World Wide Web and become more connected to the global economy. In many developing nations, the cost of online access has meant that a relatively few individuals can use online media. Yet, as a whole, the aggregate of the world's poorer populations constitutes a powerful market companies can tap. The buying power of Rio de Janeiro's poorest residents, for example, is estimated to be some $1.2 billion (Beyond the Digital Divide, 2004). Based on economies of scale, poorer consumers in developing nations could constitute an important future market; all organizations need is a mechanism for interacting with these consumers.

Many businesses see online media as an important conduit for accessing these overseas markets. It is, perhaps, for this reason that certain companies have started developing online communication technologies that could provide the less well off citizens of the world with affordable online access (Beyond the Digital Divide, 2004; Kalia, 2001). They also have begun developing inexpensive hubs for online access in nations such as India, Ghana, Brazil, and South Africa (Beyond the Digital Divide, 2004). Should such activities prove profitable, then other organizations likely will follow this lead and try to move into these markets. Thus, the international growth in OSS use has great prospects to offer a variety of organizations.

The problem with making these business situations a reality has to do with compatibility. That is, if businesses use proprietary software to create materials that are then sent to outsourced employees who work with OSS, can those products be used? Conversely, could the material produced by outsource workers be used by the client company or even by the consumers to which those materials will be marketed? Further, as OSS allows individuals to personalize the software they use, each overseas employee, in theory, could be working with a different program. Companies that then collect components created by numerous outsourcing workers thus could find the task of assembling the desired final product tedious, if not impossible. This degree of individualization also could mean that prospective consumers in developing nations use a variety of programs to access online materials, a factor that could make the task of mass marketing in these regions highly difficult.

For these reasons, future outsourcing and international marketing operations likely will require protocols and systems of standards, if individuals wish to maximize the potential of both situations. Fortunately, OSS use in developing nations is still relatively limited, so organizations do have time to work on viable protocols and standards now, at a time when the adoption of such standards would be relatively easy. Thus, an understanding of OSS no longer can be viewed as a novelty or an interesting alternative; instead, it needs to be seen as a requirement for organizational success in the future. By understanding the benefits and limitations of OSS,

individuals can make informed decisions about how to establish international protocols for working with OSS.

CONCLUSION

New technologies bring with them new choices. While Open Source software has existed for some time, it is still viewed as new by many organizations. Recent international business developments, however, indicate that the importance of OSS will grow markedly in the future. To prepare for such growth, individuals need to understand what OSS is, the benefits it has to offer, and the limitations that affect its use. By expanding their knowledge of OSS, employees and managers can better prepare themselves for the workplaces of the future.

REFERENCES

Beyond the digital divide. (2004, March 13). *The Economist: Technology Quarterly Supplement*, 8.

Brase, R. (2003, March 19). Open source makes business sense. Retrieved April 22, 2003, from *http://www.zdnet.com.au/newstech/os/story/0,2000048630,20272976,00.htm*

China's economic power. (2001, March 10). The Economist, 23-25.

Clifford, M., & Roberts, D. (2001, April 16). China: Coping with its new power. *BusinessWeek,* 28-34.

Cox, A. (2003). The risks of closed source computing. Retrieved April 22, 2003, from *http://www.osopinion.com/Opinions/AlanCox/AlanCox1.html*

DiBona, C., Ockman, S., & Stone, M. (1999). Introduction. In C. DiBona, S. Ockman, & M. Stone (Eds.), *Open sources: Voices of the open source revolution* (pp. 1-17). Sebastopol, CA: O'Reilly & Associates, Inc.

Goth, G. (2001). The open market woos open source. *IEEE Software, 18*(2), 104-107.

Kalia, K. (2001, July/August). Bridging global digital divides. *Silicon Alley Reporter,* 52-54.

Kenwood, C.A. (2001). A business case study of open source software. Bedford, MA: The MITRE Corporation.

Koch, C. (2003). Open source—Your open source plan. *CIO, 16*(11), 52-59.

Liu, E. (2002, June 10). Governments embrace open source. Retrieved May 27, 2003, from *http://www.osopinion.com*

Malik, R. (2004, July). The new land of opportunity. *Business 2.0,* 72-79.

Moody, G. (1998). The wild bunch. *New Scientist, 160*(2164), 42-46.

Nadan, C.H. (2002, Spring). Open source licensing: Virus or virtue? [electronic vVersion]. *Texas Intellectual Property Law, 10*(3), 349-377. Retrieved July 1, 2003, from *http://firstsearch.oclc.org/FSIP?sici=1068-1000%28200221 %2910%3A3%3C349%3AOSLVOV%3E&dbname=Wilson SelectPlus_FT*

Open source's local heroes. (2003, December 6). *The Economist: Technology Quarterly Supplement,* 3-5.

Perens, B. (1999). The open source definition. In C. DiBona, S. Ockman, & M. Stone (Eds.), *Open sources: Voices of the open source revolution* (pp. 171-188). Sebastopol, CA: O'Reilly & Associates, Inc.

Raymond, E.S. (1999). *The cathedral and the bazaar: Musings on Linux and open source by an accidental revolutionary.* Sebastopol, CA: O'Reilly & Associates, Inc.

KEY TERMS

Copyleft: A term coined by Richard Stallman, leader of the free software movement and creator of the General Public License, or GPL. The key tenet of the GPL, which copyleft describes, is that software licensed under it can be freely copied, distributed, and modified. Hence, this software is copyleft, or the opposite of copyright. It insures that there are no protections or restrictions when copyright insures the opposite.

Forking: Source code is said to fork when another group of developers create a derivative version of the source code that is separate, if not incompatible, with the current road the source code's development is following. The result is source code that takes a different fork in the road.

International Outsourcing: A production model in which online media are used to send work to employees located in a nation (generally, a developing nation)

Open Source Software: In general, software available to the general public to use or modify free of charge is considered open. It is also considered open source because it is software that is typically created in a collaborative environment in which developers contribute and share their programming openly with others.

Proprietary Software: Software, including the source code, that is privately owned and controlled.

Source Code: Programmers write software in source code, which are instructions for the computer to tell it how to use the software. But computers need a machine language to understand, so the source code of the software must be compiled into an understandable object code that computers can use to carry out the software instructions or source code. Without source code, a software's instructions or functionality cannot be modified. Source code that can be accessed by the general public is considered open (open source software). If it cannot be accessed, it is considered closed (proprietary software).

Optical Burst Switching

Joel J.P.C. Rodrigues
Universidade de Beira Interior, Portugal

Mário M. Freire
Universidade de Beira Interior, Portugal

Paulo P. Monteiro
SIEMENS S.A. and Universidade de Aveiro, Portugal

Pascal Lorenz
University of Haute Alsace, France

INTRODUCTION

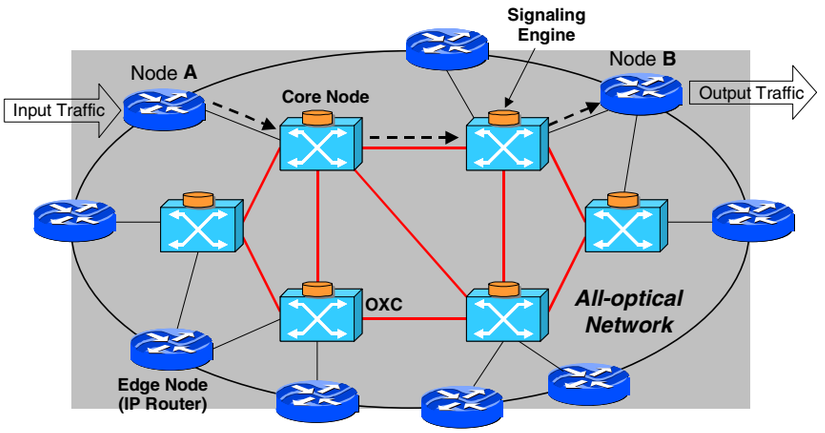
The concept of burst switching was proposed initially in the context of voice communications by Haselton (1983) and Amstutz (1983; 1989) in the early 1980s. More recently, in the late 1990s, optical burst switching (OBS) was proposed as a new switching paradigm for the so-called optical Internet in order to overcome the technical limitations of optical packet switching; namely, the lack of optical random access memory (optical RAM) and to the problems with synchronization.(Yoo & Qiao, 1997; Qiao & Yoo, 1999; Chen, Qiao & Yu, 2004; Turner, 1999; Baldine, Rouskas, Perros & Stevenson, 2002; Xu, Perros & Rouskas, 2001). OBS is a technical compromise between wavelength routing and optical packet switching, since it does not require optical buffering or packet-level processing as in optical packet switching, and it is more efficient than circuit switching if the traffic volume does not require a full wavelength channel. According to Dolzer, Gauger, Späth, and Bodamer (2001), OBS has the following characteristics:

- **Granularity:** The transmission unit size (burst) of OBS is between the optical circuit switching and optical packet switching.
- **Separation Between Control and Data:** Control information (header) and data are transmitted on different wavelengths (or channels) with some time interval.
- **Allocation of Resources:** Resources are allocated using mainly one-way reservation schemes. A source node does not need to wait for the acknowledge message from destination node to start transmitting the burst.
- **Variable Burst Length:** The burst size is variable.
- **No Optical Buffering:** Burst switching does not require optical buffering at the intermediate nodes (without any delay).

In OBS networks, IP packets (datagrams) are assembled into very large size packets called data bursts. These bursts are transmitted after a burst header packet (also called setup message or control packet) with a delay of some offset time in a given data channel. The burst offset is the interval of time at the source node between the processing of the first bit of the setup message and the transmission of the first bit of the data burst. Each control packet contains routing and scheduling information and is processed in core routers at the electronic level before the arrival of the corresponding data burst (Baldine, Rouskas, Perros & Stevenson, 2002; Qiao & Yoo, 1999; Verma, Chaskar & Ravikanth, 2000; White, Zukerman & Vu, 2002). The transmission of control packets forms a control network that controls the routing of data bursts in the optical network (Xiong, Vandenhoute & Cankaya, 2000). Details about OBS network architecture are given in the next section.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Schematic representation of an OBS network



OBS NETWORK ARCHITECTURE

An OBS network is an all-optical network where core nodes, composed by optical cross connects (OXC) plus signaling engines, transport data from/to edge nodes (IP routers), being the nodes interconnected by bi-directional links, as shown in Figure 1. This figure also shows an example of an OBS connection, where input packets come from the source edge node A to the destination edge node B. The source edge node is referred to as the ingress node, and the destination edge node is referred to as the egress node. The ingress node of the network collects the upper layer traffic and sorts and schedules it into electronic input buffers, based on each class of packets and destination address. These packets are aggregated into bursts and are stored in the output buffer, where electronic RAM is cheap and abundant (Chen, Qiao & Yu, 2004). After the burst assembly process, the control packet is created and immediately sent toward the destination to set up a connection for its corresponding burst. After the offset time, bursts are all optically transmitted over OBS core nodes without any storage at the intermediate nodes within the core, until the egress node. At the egress node, after the reception of a burst, the burst is disassembled into IP packets and provides these IP packets to the upper layer. These IP packets are forwarded electronically to destination users (Kan, Balt, Michel & Verchere, 2002; Vokkarane, Haridoss & Jue, 2002; Vokkarane & Jue, 2003).

OBS EDGE NODES

The OBS edge node works like an interface between the common IP router and the OBS backbone (Kan, Balt, Michel & Verchere, 2002; Xu, Perros & Rouskas, 2003). An OBS edge node needs to perform the following operations:

- Assemblies IP packets into data bursts based on some assembly policy;
- Generates and schedules the control packet for each burst;
- Converts the traffic destined to the OBS network from the electronic domain to the optical domain and multiplexes it into the wavelength domain;
- Demultiplexes the incoming wavelength channels and performs optical-to-electronic conversion of the incoming traffic;
- Disassembles and forwards IP packets to client IP routers.

The architecture of the edge node includes three modules (Vokkarane & Jue, 2003): routing module, burst assembly module, and scheduler module. Figure 2 shows the architecture of an edge node (Chen, Qiao & Yu, 2004; Vokkarane, Haridoss & Jue, 2002; Vokkarane & Jue, 2003). The routing module selects the appropriate output port for each packet and sends it to the correspondent burst assembly module. The burst assembly module assembles bursts containing packets that are addressed for a specific

Figure 2. Architecture of an OBS edge node

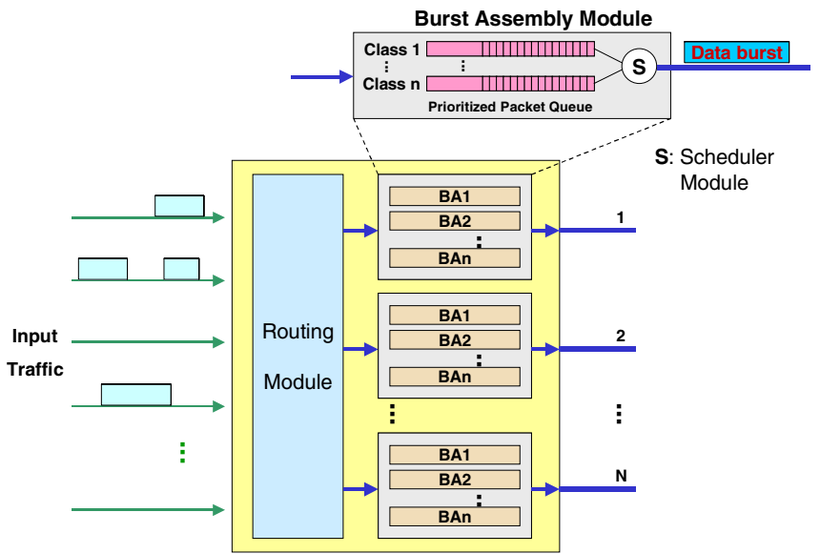
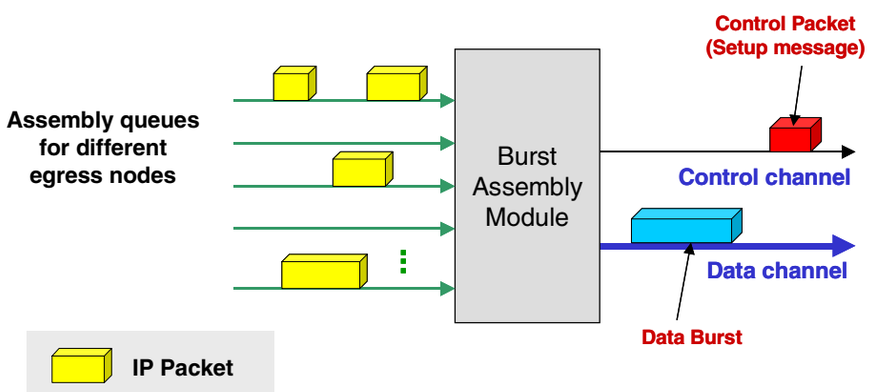


Figure 3. Burst assembly process



egress node. In this module, there is a different packet queue for each class of traffic. Usually, there are different assembly queues for each class of traffic (or priority). The burst scheduler module creates a burst and their corresponding control packet, based on the burst assembly policy, and sends it to the output port.

Burst assembly process is the most important task performed by the edge node. In OBS networks, burst assembly (Cao, Li, Xen & Qiao, 2002; Vokkarane & Jue, 2003; Xiong, Vandenhoute & Cankaya, 2000) is basically the process of aggregating and assembling packets into bursts at the ingress edge node. At this node, packets that are destined for the same egress node and belong to the same Quality of Service (QoS) class are aggregated and sent in discrete bursts, with times determined by the burst assembly policy. The burst assembly process is made into the burst assembly module inside the edge node (see Figure 3). Packets that are destined to different egress nodes go through different assembly queues to burst assembly module. In this module, the data burst is assembled, and the corresponding burst control packet is generated. At the egress node, the burst is subsequently deaggregated and forwarded electronically.

In the burst assembly process, there are two parameters that determine how the packets are

aggregated: the maximum waiting time (timer value) and the minimum size of the burst (threshold value). Based on these parameters, some burst assembly algorithms have been proposed:

- Timer-based algorithm (Cao, Li, Xen & Qiao, 2002)
- Burst length-based algorithm (Vokkarane, Haridoss & Jue, 2002)
- Hybrid algorithm or mixed timer/threshold-based algorithm (Chen, Qiao & Yu, 2004; Xiong, Vandenhoute & Cankaya, 2000).

Recently, a mechanism was proposed to provide QoS support that considers bursts containing a combination of packets with different classes, called composite burst assembly (Vokkarane & Jue, 2003). This mechanism was proposed to make good use of burst segmentation, which is a technique used for contention resolution in the optical core network, where packets toward the tail of the burst have a larger probability of being dropped than packets at the head of the burst. The authors concluded that approaches with composite bursts perform better than approaches with single-class bursts in terms of burst loss and delay, providing differentiated QoS for different classes of packets.

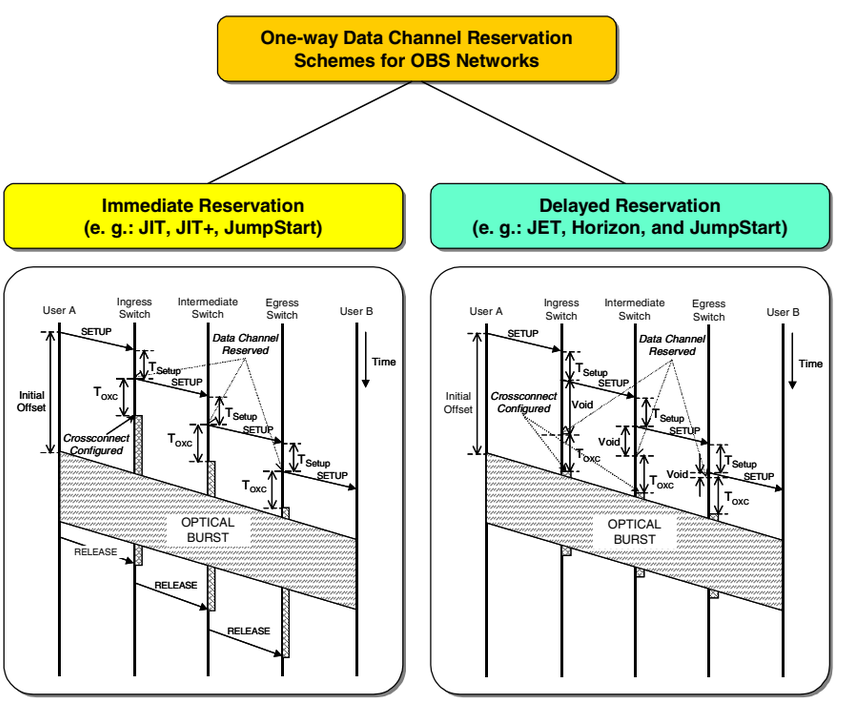
OBS CORE NODES

An OBS core node consists of two main components (Vokkarane & Jue, 2003; Xiong, Vandenhoute & Cankaya, 2000): an optical cross connect (OXC) and a switch control unit (SCU), also called a signaling engine. The SCU implements the OBS signaling protocol and creates and maintains the forwarding table and configures the OXC.

Kan, Balt, Michel, and Verchere (2001) summarize the operations that an OBS core node needs to perform, which are the following:

- Demultiplexes the wavelength data channels;
- Terminates data burst channels and conducts wavelength conversion for passing through the optical switch fabric;
- Terminates control packets channels and converts the control information from the optical to electronic domain;

Figure 4. Classification of one-way reservation schemes for optical burst switching networks



- Schedules the incoming bursts, sends the instructions to the optical switch matrix, and switches burst channels through the optical switch matrix;
- Regenerates new control packet for outgoing bursts;
- Multiplexes outgoing control packets and bursts together into single or multiple fibers.

Signaling is an important issue in the OBS network architecture, because it specifies the protocol that OBS nodes use to communicate connection requests to the network. The operation of this signaling protocol determines whether or not the resources are utilized efficiently. According to the length of the burst offset, signaling protocols may be classified into three classes: no reservation, one-way reservation, and two-way reservation (Xu, 2002).

In the first class, the burst is sent immediately after the setup message, and the offset is only the transmission time of the setup message. This first class is practical only when the switch configuration time and the switch processing time of a setup message are very short. The Tell-and-Go (TAG) protocol (Widjaja, 1995) belongs to this class. In signaling protocols with one-way reservation, a burst is sent shortly after the setup message, and the source node does not wait for the acknowledgement sent by the destination node. Therefore, the size of the offset is between the transmission time of the setup message and the round-trip delay of the setup message. Different optical burst switching mechanisms may choose different offset values in this range. JIT (Just-In-Time) (Wei & McFarland, 2000), JIT+ (Teng & Rouskas, 2005), JumpStart (Baldine, Rouskas, Perros & Stevenson, 2002; Baldine, Rouskas, Perros & Stevenson, 2003; Zaim, Baldine, Cassada, Rouskas, Perros & Stevenson, 2003), JET (Just-Enough-Time) (Qiao & Yoo, 1999), and Horizon (Turner, 1999) are examples of signaling protocols using one-way reservation schemes. The offset in two-way reservation class is the interval of time between the transmission of the setup message and the reception of the acknowledgement from the destination. The major drawback of this class is the long offset time, which causes a long data delay. Examples of signaling protocols using this class include the Tell-and-Wait (TAW) protocol (Widjaja, 1995) and the scheme proposed by Duser and Bayvel (2002). Due to the impairments of no reservation and two-way reservation classes, one-way reservation schemes seem to be more suitable for OBS networks. Therefore, the remainder of this article provides an overview of signaling protocols with one-way wavelength reservation schemes. As shown in Figure 4, one-way reservation schemes may be classified regarding the way in which output wavelengths are reserved for bursts, as immediate and delayed reservation (Rodrigues, Freire & Lorenz, 2004). JIT and JIT+ are examples of immediate wavelength reservation, while JET and Horizon are examples of delayed reservation schemes. The JumpStart signaling protocol may be implemented using either immediate or delayed reservation.

The JIT signaling protocol considers that an output wavelength is reserved for a burst immediately after the arrival of the corresponding setup message. If a wavelength cannot be reserved immediately, then the setup message is rejected, and the corresponding burst is dropped. Figure 4 illustrates the operation of JIT protocol. As may be seen in this figure, TSetup represents the amount of time that is needed to process the setup message in an OBS node, and TOXC represents the delay incurred from the instant that the OXC receives a command from the signaling engine to set up a connection from an input port to an output port until the instant the appropriate path within the optical switch is complete and can be used to switch a burst (Teng & Rouskas, 2005). JIT+ is a modified version of the immediate reservation scheme of JIT. Under JIT+, an output wavelength is reserved for a burst if (i) the arrival time of the burst is later than the time horizon of the wavelength and (ii) the wavelength has, at most, one other reservation (Teng & Rouskas, 2005). This protocol does not perform any void filling. Comparing JIT+ with JET and Horizon, the latter ones permit an unlimited number of delayed reservations per wavelength, whereas JIT+ limits the number of such operations to, at most, one per wavelength. On the other hand, JIT+ maintains all the advantages of JIT in terms of simplicity of hardware implementation.

Delayed reservation schemes, exemplified by the schemes used in JET and Horizon signaling protocols, considers that an output wavelength is reserved for a burst just before the arrival of the first bit of the burst. If, upon arrival of the setup message,

no wavelength can be reserved at a suitable time, then the setup message is rejected, and the corresponding burst is dropped (Teng & Rouskas, 2005). In these kinds of reservation schemes, when a burst is accept by the OBS node, the output wavelength is reserved for an amount of time equal to the length of the burst plus *TOXC*, in order to account for the OXC configuration time. As one may see in Figure 4, a void is created on the output wavelength between time $t + TSetup$, when the reservation operation for the upcoming burst is completed, and time $t' = t + Toffset - TOXC$, when the output wavelength actually is reserved for the burst. The Horizon protocol is an example of signaling protocols with delayed reservation scheme without void filling, and it is less complex than the signaling protocols with delayed reservation schemes with void filling, such as JET. When Horizon protocol is used, an output wavelength is reserved for a burst, if and only if the arrival time of the burst is later than the time horizon of the wavelength. If, upon arrival of the setup message, it is verified that the arrival time of the burst is earlier than the smallest time horizon of any available wavelength, then the setup message is rejected, and the corresponding burst is dropped (Teng & Rouskas, 2005).

On the other hand, JET is the most well-known signaling protocol with delayed wavelength reservation scheme with void filling, which uses information to predict the start and the end of the burst. In this protocol, an output wavelength is reserved for a burst, if the arrival time of the burst (i) is later than the time horizon of the wavelength, or (ii) coincides with a void on the wavelength, and the end of the burst (plus the OXC configuration time) occurs before the end of the void. If, upon arrival of the setup message, it is determined that none of these conditions is satisfied for any wavelength, then the setup message is rejected, and the corresponding burst is dropped (Teng & Rouskas, 2004).

Recently, it was observed by Rodrigues, Freire, and Lorenz (2004) that the above five signaling protocols lead to a similar network performance, and, therefore, the simplest protocols (i.e., JIT-based protocols) should be considered for implementation in practical systems.

CONCLUSION

OBS has been proposed to overcome the technical limitations of optical packet switching. In this article, an overview of OBS networks was presented. The overview focused on the following issues: OBS network architecture, architecture of edge nodes, burst assembly process in edge nodes, and architecture of core nodes and signaling protocols.

REFERENCES

Amstutz, S.R. (1983). Burst switching—An introduction. *IEEE Communications Magazine, 21*(8), 36-42.

Amstutz, S.R. (1989). Burst switching—An update. *IEEE Communications Magazine, 27*(9), 50-57.

Baldine, I., Rouskas, G., Perros, H., & Stevenson, D. (2002). JumpStart—A just-in-time signaling architecture for WDM burst-switched networks. *IEEE Communications Magazine, 40*(2), 82-89.

Baldine, I., Rouskas, G.N., Perros, H.G., & Stevenson, D. (2003). Signaling support for multicast and QoS within the JumpStart WDM burst switching architecture. *Optical Networks, 4*(6), 68-80.

Cao, X., Li, J., Xen, Y., & Qiao, C. (2002). Assembling TCP/IP packets in optical burst switched networks. In *Proceedings of Global Telecommunications Conference (GLOBECOM '2002).* IEEE, 3, 2808-2812.

Chen, Y., Qiao, C., & Yu, X. (2004). Optical burst switching: A new area in optical networking research. *IEEE Network, 18*(3), 16-23.

Dolzer, K., Gauger, C., Späth, J., & Bodamer, S. (2001). Evaluation of reservation mechanisms for optical burst switching. *AEÜ International Journal of Electronics and Communications, 55*(1).

Duser M., & Bayvel, P. (2002). Analysis of a dynamically wavelength-routed optical burst switched network architecture. *Journal of Lightwave Technology, 20*(4), 574-585.

Haselton, E.F. (1983). A PCM frame switching concept leading to burst switching network architecture. *IEEE Communications Magazine, 21*(6), 13-19.

Kan, C., Balt, H., Michel, S., & Verchere, D. (2001). Network-element view information model for an optical burst core switch. *Proceedings of the Asia-Pacific Optical and Wireless Communications Conference (APOC)*, Beijing, China, SPIE, 4584, 115-125.

Kan, C., Balt, H., Michel, S., & Verchere, D. (2002). Information model of an optical burst edge switch. *Proceedings of the IEEE International Conference on Communications (ICC 2002)I,* New York, New York, USA.

Perros, H. (2001). *An introduction to ATM networks*. New York: Wiley.

Qiao, C., & Yoo, M. (1999). Optical burst switching (OBS)—A new paradigm for an optical Internet. *Journal of High Speed Networks, 8*(1), 69-84.

Rodrigues, J.J.P.C., Freire, M.M., & Lorenz, P. (2004). Performance Assessment of signaling protocols with one-way reservation schemes for optical burst switched networks. In Z. Mammeri, & P. Lorenz (Eds.), *High-speed networks and multimedia communications* (pp. 821-831). Berlin: Springer-Verlag.

Teng J., & Rouskas, G.N. (2005). *A detailed analysis and performance comparison of wavelength reservation schemes for optical burst switched networks [to be published], 9*(3), 311-335. Photonic Network Communications.

Turner, J.S. (1999). Terabit burst switching. *Journal of High Speed Networks, 8*(1), 3-16.

Verma, S., Chaskar, H., & Ravikanth, R. (2000). Optical burst switching: A viable solution for Terabit IP backbone. *IEEE Network, 14*(6), 48-53.

Vokkarane V., & Jue, J. (2003). Prioritized burst segmentation and composite burst assembly techniques for QoS support in optical burst-switched networks. *IEEE Journal on Selected Areas in Communications, 21*(7), 1198-1209.

Vokkarane, V.M., Haridoss, K., & Jue, J.P. (2002). Threshold-based burst assembly policies for QoS support in optical burst-switched networks. *Proceedings of the SPIE Optical Networking and Communication Conference (OptiComm)*, Boston, Massachusetts, USA.

Wei, J.Y. & McFarland, R.I. (2000). Just-in-time signaling for WDM optical burst switching networks. *Journal of Lightwave Technology, 1*8(12), 2019-2037.

White, J., Zukerman, M., & Vu, H.L. (2002). A framework for optical burst switching network design. *IEEE Communications Letters, 6*(6), 268-270.

Widjaja, I. (1995). Performance analysis of burst admission control protocols. *IEE Proceeding of Communication, 142*, 7-14.

Xiong, Y., Vandenhoute, M., & Cankaya, H.C. (2000). Control architecture in optical burst-switched WDM networks. *IEEE Journal on Selected Areas in Communications, 18*(10), 1838-1851.

Xu, L. (2002). Performance analysis of optical burst switched networks [Ph.D. thesis]. Raleigh, NC: North Carolina State University.

Xu, L., Perros, H.G., & Rouskas, G.N. (2001). Techniques for optical packet switching and optical burst switching. *IEEE Communications Magazine, 39*(1), 136-142.

Xu, L., Perros, H.G., & Rouskas, G.N. (2003). A queueing network model of an edge optical burst switching node. *IEEE INFOCOM, 3*, 2019-2029.

Yoo, M., & Qiao, C. (1997). Just-enough-time (JET): A high speed protocol for bursty traffic in optical networks. *Proceedings of the IEEE/LEOS Conference on Technologies for a Global Information Infrastructure*, Montreal, Quebec, Canada.

Zaim, A.H., et al. (2003). The JumpStart just-in-time signaling protocol: A formal description using EFSM. *Optical Engineering, 42*(2), 568-585.

KEY TERMS

Burst Assembly: Basically the process of aggregating and assembling packets into bursts at the ingress edge node of an OBS network.

Burst Offset: The interval of time at the source node between the processing of the first bit of the setup message and the transmission of the first bit of the data burst.

Bursts: In OBS networks, IP packets (datagrams) are assembled into very large sized data packets called bursts.

Control Packet (or Bburst Header Packet or Setup Message): A control packet is sent in a separated channel and contains routing and scheduling information to be processed at the electronic level before the arrival of the corresponding data burst.

Network Architecture: Defines the structure and the behavior of the real subsystem that is visible for other interconnected systems, while they are involved in the processing and transfer of information sets.

One-Way Reservation Schemes: These schemes may be classified, regarding the way in which output wavelengths are reserved for bursts, as immediate and delayed reservation. JIT and JIT+ are examples of immediate wavelength reservations, while JET and Horizon are examples of delayed reservation schemes.

Optical Cross-Connect (OXC): Optical device used mainly in long-distance networks that can shift signals from an incoming wavelength to an output wavelength of a given optical fiber.

Quality of Service (QoS): Represents a guarantee or a commitment not only to a particular quality of network service, but also a particular rate or minimum rate of data delivery, as well as maximum transmission times among packets.

SCU: Switch control unit or signaling engine. The SCU implements the OBS signaling protocol, creates and maintains the forwarding table, and configures the optical cross connect.

Peer-to-Peer Filesharing Systems for Digital Media

Jerald Hughes
Baruch College of the City University of New York, USA

Karl Reiner Lang
Baruch College of the City University of New York, USA

INTRODUCTION

In 1999, exchanges of digital media objects, especially files of music, came to constitute a significant portion of Internet traffic thanks to a new set of technologies known as peer-to-peer (P2P) file-sharing systems. The networks created by software applications such as Napster and Kazaa have made it possible for millions of users to gain access to an extraordinary range of multimedia files, which, by virtue of their purely digital form, have the desirable characteristics of portability and replicability, which pose great challenges for businesses that have in the past controlled images and sound recordings.

Peer-to-peer is a type of network architecture in which various nodes have the capability of communicating directly with other nodes without having to pass messages through any central controlling node (Whinston, Parameswaran, & Susarla, 2001). The basic infrastructure of the Internet relies on this principle for fault tolerance; if any single node ceases to operate, messages can still reach their destination by rerouting through other still-functioning nodes. The Internet today consists of a complex mixture of peer-to-peer and client-server relationships, but P2P file-sharing systems operate as overlay networks (Gummadi, Saroiu, & Gribble, 2002) upon that basic Internet structure.

P2P file-sharing systems are software applications that allow direct communications between nodes in the network. They share this definition with other systems used for purposes other than file sharing, such as instant messaging, distributed computing, and media streaming. What these P2P technologies have in common is the ability to leverage the combined power of many machines in a network to achieve results that are difficult or impossible for single machines to accomplish. However, such networks also open up possibilities for pooling the interests and actions of the users so that effects emerge that were not necessarily anticipated when the network technology was originally created (Castells, 2000).

TECHNICAL FOUNDATIONS

In order for P2P file-sharing systems to function, several digital technologies had to come together (see Table 1).

Digital media files are large, and until both low-cost broadband connections and effective compression technologies became available, the distribution of digital media objects such as popular songs was not practical. Today, with relatively affordable broadband Internet access widely available in much of the world, anyone who wishes to use a P2P file-sharing application can do so.

The first digital format for a consumer product was the music CD (compact disc), introduced in the early 1980s. This format, known as Redbook Audio, encoded stereo sound files using a sample rate of 44.1 kHz and a sample bit depth of 16 bits. In Redbook Audio, a song 4 minutes long requires 42 Mb of storage. A music CD, with roughly 700 Mb of storage, can thus hold a little over an hour of music. Even at broadband speeds, downloading files of this size is impractical for many users, so the next necessary component of file sharing is effective compression.

The breakthrough for file sharing came from the MPEG specification for digital video via the Fraunhofer Institute in Erlangen, Germany, which examined the MPEG-1 Layer 3 specification and developed the first stand-alone encoding algorithms for MP3 files. Layer 3 deals with the audio tracks of

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. Enabling technologies for P2P file-sharing systems

Broadband Internet access	T1 and T3 digital transmission lines, DSL, cable modems, satellite
Encoding for digital media	Music: MP3 (MPEG [Motion Picture Experts Group] 1 Layer 3), Advanced Audio Coding (AAC), Windows Media Audio (WMA) Movies and video: Digital Video Express (DivX) (MPEG 4)
Multimedia players	Software: Winamp, MusicMatch, RealPlayer Hardware: iPod, Rio
P2P overlay networks	Napster, Kazaa, BitTorrent, Grokster, Limewire

a video recording. The MP3 encoding algorithm makes use of a psychoacoustic phenomenon known as masking to discard portions of the sound spectrum that are unlikely to be heard during playback, yielding a compression ratio for standard MP3 files of 11:1 from the original Redbook file. Standard MP3 encoding uses a bit-stream rate of 128 Kb per second, although MP3 encoding tools now allow variable rates. With a single song of 4 minutes in length available in relatively high-quality form in a digital file *only* 4 Mb large, the stage was set for the emergence of P2P file-sharing applications. The killer app for MP3 users was Winamp, a free software application able to decode and play MP3 files. The widespread adoption of the MP3 format has made it necessary for developers of other media applications such as Windows Media Player and RealPlayer to add MP3 playback capabilities to their media platforms. P2P applications can make any file type at all available; while MP3s are the most popular for music, many other file types also appear, including .wav (for audio), .exe (computer programs), .zip (compressed files), and many different formats for video and images.

In order to get an MP3 file for one's Winamp, it is necessary to either make it oneself from a CD (ripping), or find it on a file-sharing network. Applications such as Napster and Kazaa use metadata to allow keyword searches. In the original Napster, keywords went to a central server that stored an index of all files in the system, and then gave the file seeker the IP (Internet protocol) address of a machine, the servant, which contained a file whose metadata matched the query. This system thus used centralized index and distributed storage. The Gnutella engine, which is the basis for P2P systems such as Kazaa, Limewire, and others, uses a more purely peer-to-peer architecture in which no central index is required (Vaucher, Babin, Kropf, & Jouve, 2002). A Gnutella query enters the P2P network looking for keyword matches on individual computers rather than in a central index. This architectural difference means that a Kazaa search may be less complete than a Napster search because Gnutella queries include a "time-to-live" (TTL) attribute that terminates the query after it crosses seven network nodes.

Once a desired file is discovered, the P2P application establishes a direct link between the two machines so the file can be downloaded. For Napster, Kazaa, and many other P2P systems, this involves a single continuous download from a single IP address. In order to handle multiple requests, queues are set up. Users may share all, none, or some of the files on their hard drives. Sharing no files at all, a behavior known as free riding (Adar & Huberman, 2000), can degrade the performance of the network, but the effect is surprisingly small until a large majority of users are free riding. For individual song downloads, using one-to-one download protocols works well, but for very large files, such as those used for digital movie files, the download can take hours and may fail entirely if the servant leaves the network before transfer is complete. To solve this problem, the P2P tool BitTorrent allows the user to download a single file from many users at once, thus leveraging not only the storage but also the bandwidth of the machines on the network. A similar

technique is used by P2P systems that provide streaming media (not file downloads) in order to avoid the cost and limitations of single-server media delivery.

APPLICATIONS

P2P file-sharing systems have already passed through at least three stages of development. The original Napster was closed by the courts because the system's use of a central index was found to constitute support of copyright infringement by users. The second generation, widely used in tools such as Kazaa, reworked the architecture to allow for effective file discovery without the use of a central index; this system had withstood legal challenges as of 2004. However, the users themselves were exposed to legal sanctions as the Recording Industry Association of America (RIAA) filed lawsuits on behalf of its members against users who made files of music under copyright available to other network users. The third generation of P2P file-sharing tools involves a variety of added capabilities, including user anonymity, more efficient searches, and the ability to share very large files. One prominent third-generation architecture is Freenet (Clarke, Sandberg, Wiley, & Hong, 2001), which provides some, but not perfect, anonymity through a protocol that obscures the source of data requests. Krishnan and Uhlmann (2004) have designed an architecture that provides user anonymity by making file requests on behalf of a large pool of users, thus providing a legal basis for plausible deniability for any particular file request.

There are now dozens of P2P file-sharing applications available, sharing every conceivable type of media object in every available format. Video presented a difficult problem for file sharing until the development of the MPEG-4 specification, which allows one to create high-quality movie-length video files in *only* hundreds of megabytes instead of gigabytes of data. DivX is a format derived from MPEG-4. P2P networks can also be a medium of exchange for digital images. Müller and Henrich (2003) have presented a P2P file-sharing architecture that, instead of relying on keywords, would allow users to search for images based on image feature vectors that represent, for example, color, texture, or shape properties.

Information-retrieval techniques can be applied to P2P file-sharing networks in searches for text documents. Lu and Callan (2003) have developed a method for P2P networks that provides results based on the actual similarity of text content rather than on document names. As P2P networks continue to grow in the diversity of file types available, and especially as they are adopted for uses by institutions and businesses, text-based retrieval methods are likely to increase in importance.

Considerable work has been done to explore the usefulness of P2P networks in supporting the delivery of streamed digital media content. While the P2P networks add enormous power in the form of computational and bandwidth resources, they also have unpredictable elements, such as the fact that peers in the network may enter or leave the network at any time, which needs to be taken into account in attempting to implement content delivery with a known quality of service. Hefeeda, Habib, Botev, Xu, and Bhargava (2003) have created a P2P architecture that solves this problem by taking network topology and reliability of peers into account, and by dynamically switching in new file senders as existing ones drop out so that the resulting performance of the network remains satisfactory.

ECONOMIC IMPLICATIONS

While the major record companies have directly blamed the P2P file-sharing tools for a slump in the sales of CDs, the economic impacts of Napster and its descendants are not at all clear. There may be other reasons for a decline in CD sales; a study that used rigorous empirical research to analyze patterns of CD sales and downloads of the same music on P2P systems was unable to detect any significant effect of P2P systems on sales (Oberholzer & Strumpf, 2004). While the RIAA has tried, at least through 2004, to prevent all downloads of copyrighted digital media files, this tactic may not be in the record companies' best economic interest. Some analyses indicate that the profit-maximizing strategy could include a certain background level of music file sharing (Bhattacharjee, Gopal, & Sanders, 2003). This makes sense if one considers the

fact that the huge variety of music on such systems could make them, in effect, a marketing tool for some users who might first discover artists via the P2P systems, then purchase CDs of music they would otherwise not have considered.

P2P file-sharing systems turn traditional economic concepts upside down. Pricing mechanisms ordinarily depend upon supply-and-demand relationships based upon the assumption of limited resources. Digital media products, such as music CDs, conformed to this assumption as long as the information was locked into a physical storage device. However, pure information multimedia objects have no natural restraints on supply. Digital media files can be replicated at essentially zero cost. Furthermore, the collective effect of the actions of users on a P2P file-sharing network reverses the normal economic equation. For physical products such as oil, cars, and so forth, the more the resource is consumed, the less there is of it to go around. On P2P file-sharing systems, precisely the opposite is true: The more demand there is for a file, the greater the supply and the easier it is to acquire. P2P systems thus turn the economics of scarcity into the economics of abundance (Hughes & Lang, 2003). P2P systems also have a built-in natural resistance to attempts to degrade them. One response of the music industry to P2P file-sharing systems has been to introduce fake music files into the network (known as spoofing) in an attempt to degrade the usefulness of the system. However, the P2P system is naturally self-cleansing in this respect: The files people do want become easily available because they propagate through the network in a spreading series of uploads and downloads, while undesirable files are purged from users' hard drives and thus typically fail

Table 2. Future impacts of P2P file-sharing systems

Status Quo	***Future Trends***
Centralized markets — five major record labels seeking millions of buyers for relatively homogeneous product, media market concentration, economies of scale	***Niche Markets —*** *thousands of music producers catering to highly specific tastes of smaller groups of users, market fragmentation, economies of specialization*
Planned, rational — corporate marketing decisions based on competitive strategies	***Self-organizing, emergent –****based on the collaborative and collective actions of millions of network users (digital community networks)*
Artifact based — CD, SuperAudio CD, DVD-Audio (DVD-A)	***Information based —*** *MP3, .wav, RealAudio*
Economics of scarcity — supply regulated by record labels, physical production and distribution	***Economics of abundance —*** *P2P networks use demand to create self-reproducing supply: the more popular a file is, the more available it becomes*
Mass distribution — traditional retail distribution channels, Business-to-Consumer (B2C) (online shopping)	***P2P distribution —*** *direct user-to-user distribution via file-sharing networks (viral marketing)*
Centralized content control — product content based on the judgment of industry experts (artist and repertoire [A&R])	***Distributed content availability —*** *determined by collective judgment of users, any content can be made available*
Product-based revenues — retail sales of packaged CDs	***Service-based revenues*** *— subscription services, creation of secondary markets in underlying IT production and playback hardware and software*
Creator/consumer dichotomy — industry (stars, labels) creates music, buyer as passive consumer of finished product	***Creator/consumer convergence —*** *user has power, via networks, to participate in musical process*

to achieve significant representation in the network. Considerable work has already been done to improve P2P protocols to take the trustworthiness of nodes into account, which would make spoofing even less effective (Kamvar, Schlosser, & Garcia-Molina, 2003).

P2P FILE-SHARING SYSTEMS AND COPYRIGHT ISSUES

Existing copyright law reserves the right to distribute a work to the copyright owner; thus, someone who makes copyrighted material available for download on a P2P network can be considered to be infringing on copyrights. However, copyright law is not solely for protecting intellectual property rights; its stated purpose in the U.S. constitution is "[t]o promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries" (Article I, Section 8). The point of copyright law is thus, ideally, to regulate access to copyrighted works for a limited period of time.

In 1998 the United States passed a law called the Digital Millenium Copyright Act (DMCA), which attempted to address this balance in regard to digital information products. The DMCA has been criticized by some as unnecessarily reducing the scope of the fair use of digital media objects (Clark, 2002). The existence of P2P file-sharing systems that so easily allow the uncontrolled dissemination of digital media objects has caused copyright owners in the entertainment industry to explore digital rights-management (DRM) technologies to preserve copyrights. Tanaka (2001) recommends a stronger reliance on DRM in view of the legal justifications for allowing P2P file-sharing systems to continue to operate. The problem is that strong DRM tends to remove rights that in predigital times were unproblematically considered fair use, such as the right to play legally acquired content on the platform of one's choosing. For example, one might play a vinyl record on any record player without restrictions, while a DRM-protected iTunes digital music file can only be played on an iTunes-licensed platform, such as an iPod. P2P systems present copyright holders with a difficult problem: To allow traditional fair-use rights for digital media files makes those files too vulnerable to piracy, while too restrictive DRM protection interferes with fair use and tends to make users unhappy with the product.

The courts have ruled, using the precedent of *Sony Corporation of America v. Universal City Studios* in 1984, that technologies that have substantial noninfringing uses cannot be deemed illegal per se. That case involved the use of videocassette recorders, but the principle described in the decision is the same: Since P2P file-sharing systems have substantial legitimate uses, the P2P file-sharing systems must be allowed to continue to operate, and legal responsibility for copyright infringements must be placed elsewhere. Napster did not have this legal protection because it maintained a central index of music files, which, with Napster's knowledge, substantially contributed to copyright-infringement activities by its users.

FUTURE TRENDS

Table 2 summarizes some of the effects that we can expect from the continuing development and adoption of P2P file-sharing systems.

CONCLUSION

After an initial period of seeming indifference, the entertainment industry has begun to evolve in response to P2P-network use by its customers. Those customers, by virtue of their participation in networks, have had dramatic impacts on the value chain of multimedia products. The P2P networks provide for easy storage and replication of the product. They also, through the collective filtering effect of millions of user choices, implement the product selection process that has traditionally been carried out by artist-and-repertoire specialists in record companies. The changes wrought by P2P file-sharing systems are, and will continue to be, deep and pervasive.

REFERENCES

Adar, E., & Huberman, B. (2000). Free riding on Gnutella. *First Monday, 5*(10). Retrieved March 8,

2005, from *http://www.firstmonday.dk/issues/issue5_10/adar/*

Bhattacharjee, S., Gopal, R., & Sanders, G. (2003). Digital music and online sharing: Software piracy 2.0? *Communications of the ACM, 46*(7), 107-111.

Castells, M. (2000). *The rise of the network society* (2nd ed.). Oxford, United Kingdom: Blackwell Publishers.

Clark, I., Sandberg, O., Wiley, B. & Hong, T. (2001). Freenet: A distributed anonymous information storage and retrieval system. *Proceedings of the Workshop on Design Issues in Anonymity,* 46-66.

Clarke, I., Sandberg, O., Wiley, B., & Hong, T. (2001). Freenet: A distributed anonymous information storage and retrieval system in designing privacy enhancing technologies. *International Workshop on Design Issues in Anonymity and Unobservability*.

Gummadi, P., Saroiu, S., & Gribble, S. (2002). A measurement study of Napster and Gnutella as examples of peer-to-peer file sharing systems. *Computer Communication Review, 32*(1), 82.

Hefeeda, M., Habib, A., Botev, B., Xu, D., & Bhargava, B. (2003). PROMISE: Peer-to-peer media streaming using CollectCast. *Proceedings of the 11th ACM International Conference on Multimedia*, 45-54.

Hughes, J., & Lang, K. (2003). If I had a song: The culture of digital community networks and its impact on the music industry. *The International Journal on Media Management 5*(3), 180-189.

Kamvar, S., Schlosser, M., & Garcia-Molina, H. (2003). The eigentrust algorithm for reputation management in P2P networks. *Proceedings of the 12th International Conference on World Wide Web*, 640-651.

Krishnan, S., & Uhlmann, J. (2004). The design of an anonymous file-sharing system based on group anonymity. *Information and Software Technology, 46*(4), 273-279.

Lu, J., & Callan, J. (2003). Content-based retrieval in hybrid peer-to-peer networks. *Proceedings of the 12th International Conference on Information and Knowledge Management*, 199-206.

Müller, W., & Henrich, A. (2003). Fast retrieval of high-dimensional feature vectors in P2P networks using compact peer data summaries. *Proceedings of the Fifth ACM SIGMM International Workshop on Multimedia Information Retrieval*, 79-86.

Oberholzer, F., & Strumpf, K. (2004). *The effect of file sharing on record sales: An empirical analysis.* Retrieved September 23, 2004, from *http://www.unc.edu/~cigar/papers/FileSharing_March2004.pdf*

Tanaka, H. (2001). Post-Napster: Peer-to-peer file sharing systems. Current and future issues on secondary liability under copyright laws in the United States and Japan. *Entertainment Law Review, 22*(1), 37-84.

Vaucher, J., Babin, G., Kropf, P., & Jouve, T. (2002). Experimenting with Gnutella communities. *Proceedings of the Conference on Distributed Communities on the Web*, 85-99.

Whinston, A., Parameswaran, M., & Susarla, A. (2001). P2P networking: An information sharing alternative. *IEEE Computer, 34*(7), 31-38.

United State Constitution. Aritlce I, Section 8. Retrieved March 8, 2005, from *http://www.house.gov/Constitution/Constitution.html*

KEY TERMS

Digital Rights Management (DRM): Technologies whose purpose is to restrict access to, and the possible uses of, digital media objects, for example, by scrambling the data on a DVD to prevent unauthorized copying.

Free Riding: Using P2P file-sharing networks to acquire files by downloading without making any files on one's own machine available to the network in return.

Killer App: A software application that is so popular that it drives the widespread adoption of a new technology. For example, desktop spreadsheet software was so effective that it made PCs (personal computers) a must-have technology for virtually all businesses.

Overlay Network: A software-enabled network that operates at the application layer of the TCP/IP (transmission-control protocol/Internet protocol).

Ripping: Converting an existing digital file to a compressed format suitable for exchange over P2P file-sharing networks, for example, converting Redbook audio to MP3 format.

Servant: A node in a P2P file-sharing network that transfers a file to a user in response to a request.

Spoofing: In P2P file-sharing networks, the practice of introducing dummy files that have the name of a popular song attached but not the actual music in order to degrade the network.

Personalized Web-Based Learning Services

Larbi Esmahi
Athabasca University, Canada

NEW TRENDS IN E-LEARNING SERVICES AND NEEDS FOR PERSONNALIZATION

New Trends

Computers have a great potential as support tools for learning; they promise the possibility of affordable, individualized learning environments. In early teaching systems, the goal was to build a clever teacher able to communicate knowledge to the individual learner. Recent and emerging work focuses on the learner exploring, designing, constructing, making sense of, and using adaptive systems as tools. Hence, the new tendency is to give the learner greater responsibility and control over all aspects of the learning process. This need for flexibility, personalization, and control results from a shift in the perception of the learning process. In fact, new trends emerging in the education domain are significantly influencing e-learning (Kay, 2001) in the following ways:

- The shift from studying in order to graduate, to studying in order to learn; most e-learners are working and have well-defined personal goals for enhancing their careers.
- The shift from student to learner; this shift has resulted in a change in strategy and control so that the learning process is becoming more cooperative than competitive.
- The shift from expertise in a domain to teaching beliefs; the classical teaching systems refer to domain and teaching expertise when dealing with the knowledge transfer process, but the new trend is based on the concept of belief. One teacher may have different beliefs from another, and the different actors in the system (students, peers, teachers), may have different beliefs about the domain and teaching methods.
- The shift from a four-year program to graduate to lifelong learning; most e-learners have a long-term learning plan related to their career needs.
- The shift to conceiving university departments as communities of scholars, but not necessarily in a single location.
- The shift to mobile learning; most e-learners are working and have little spare time. Therefore, any computer-based learning must fit into their busy schedules (at work, at home, when traveling), since they require a personal and portable system.

The One-Size-Fits-All Approach

The one-size-fits-all approach is not suitable for e-learning. This approach is not suitable for the teaching material (course content and instruction methods) or for the teaching tools (devices and interfaces). The personalization of the teaching material has been studied and evaluated in terms of the psychology of learning and teaching methods since the middle of the 20th century (Brusilovsky, 1999; Crowder, 1959; Litchfield et al., 1990; Tennyson & Rothen, 1977). The empirical evaluation of these methods showed that personalized teaching material increased the learning speed and helped learners achieve better understanding than they could have achieved with non-personalized teaching material (Brusilovsky, 2003). The personalization of teaching tools has been addressed in the context of new emerging computing environments (ubiquitous, wearable, and pervasive computing). Gallis et al. (2001) studied how medical students use various information and communication devices in the learning context and argued that " there is no 'one size fits all' device that will suite [sic] all use situations and all

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

users. The use situation for the medical students, points towards the multi-device paradigm" (Gallis et al., 2001, p. 12). The multi-device paradigm fits well with the e-learning context, in which students use different devices, depending on the situation, environment, and context.

WHAT CAN BE PERSONALIZED?

An intelligent teaching system is commonly described in terms of a four-model architecture: the interaction model, the learner's model, the domain expert, and the pedagogical expert (Wenger, 1987). The interaction model deals with the interface preferences, the presentation mode (text, image, sound, etc.), and the language. The learner model represents static beliefs about the learner and learning style and, in some cases, has been able to simulate the learner's reasoning (Paiva, 1995). The domain expert contains the knowledge about the subject matter. It deals with the domain concepts and course components (i.e., text, examples, playgrounds, etc.). The pedagogical expert contains the information on how to teach the course units to the individual learner. It consists of two main parts: teaching strategies that define the teaching rules (Vassileva, 1994) and diagnostic knowledge that defines the actions to take, depending on the learner's background, experience, interests, and cognitive abilities (Specht, 1998).

Based on these four components, individualized courses are generated and presented to the learner. Moreover, the system can adapt the instructional process on several levels:

- **Course-Content Adaptation:** Adaptive presentation by inserting, removing, sorting, or dimming fragments,
- **Course-Navigation Adaptation:** Links-adaptation support by hiding, sorting, disabling, or removing links, and by generating new links.
- **Learning Strategy:** Lecture-based learning, study-case-based learning, and problem-based learning.
- **Interfaces:** To provide the user with interfaces with the same look and feel based on his or her preferences.
- **Interaction:** To be intuitive, based on the user's profile.

ADAPTING/PERSONALIZING TO WHAT?

Most of the four components described in the previous section put user modeling in the center of any adaptation process. In fact, a teaching system's behavior can be individualized only if the system has individual models of the learners. The interaction model is almost the only component in the system that makes use of the device profile in addition to the user profile. Furthermore, in this context, we have a networked system, so the interaction model should take into consideration all the networking and connection features (i.e., bandwidth, protocol, etc.).

As we discussed in the section titled "The One-Size-Fits-All Approach," learners may use different tools depending on the situation, environment, and context.

Based on these parameters, the teaching system's adaptation can be accomplished by using three types of data:

- **User Data:** Characteristics of the user (i.e., knowledge; background; experience; preferences; user's individual traits such as personality factors, cognitive factors, and learning styles).
- **Usage Data:** Data about user interaction with the system (i.e., user's goals and tasks, user's interests).
- **Environment Data:** All aspects of the environment that are not related to the user (i.e., equipment, software, location, platform, network bandwidth).

OVERVIEW OF SOME IMPLEMENTED SYSTEMS

Since the early days of Internet expansion, researchers have implemented different kinds of adaptive and intelligent systems for Web-based education. Almost all of these systems inherited their features from the two well-known types: Intelligent

Tutoring Systems (Brusilovsky, 1995) and Adaptive Hypermedia Systems (Brusilovsky, 1996).

Intelligent tutoring research focuses on three problems: curriculum sequencing, intelligent analysis of learner's solutions, and interactive problem-solving support; whereas adaptive hypermedia systems focus on adaptive presentation and adaptive navigation support. In this section, we briefly present some implemented systems that use one or more of these concepts. For more details on these systems, the reader can refer to the cited references.

- **ELM-ART:** (Weber & Brusilovsky, 2001; Weber & Specht, 1997): An on-site intelligent learning environment that supports example-based programming, intelligent analysis of problem solutions, and advanced testing and debugging facilities. ELM-ART II supports active sequencing by using a combination of an overlay model and an episodic user model. The overlay model represents the student's problem-solving knowledge and consists of a set of goal-action or goal-plan rules. The episodic model uses a case-based approach and consists of cases describing problems and solutions selected or developed by the student. ELM-ART II also implements adaptive navigation based on the student's model. Finally, ELM-ART II supports example-based problem solving.
- **ACE:** (Specht, 2000; Specht & Oppermann, 1998): ACE is a Web-based intelligent tutoring system that combines instructional planning and adaptive media generation to deliver individualized teaching material. ACE uses three models for adapting different aspects of the instructional process: domain model, pedagogical model, and learner model. ELM-ART II was basically the starting point for ACE. Hence, ACE inherited many knowledge structures from ELM-ART II. The learner model of ACE combines a probabilistic overlay model and episodic model similar to those used in ELM-ART II. The probabilistic overlay model is used for several adaptation levels: adaptive sequencing, mastery learning, adaptive testing, and adaptive annotation. The episodic model is used to generate hypotheses about the learner's knowledge and interests. The domain model describes the domain concepts and their interrelations and dependencies. It is built on a conceptual network of learning units, where each unit can be either sections or concepts. The pedagogical model contains the teaching strategies and diagnostic knowledge. The teaching strategies define the rules for different sequencing of each concept in the learning material. The diagnostic components store the knowledge about several types of tests and how they have to be generated and evaluated. ACE supports adaptive navigation by using adaptive annotation (ELM-ART II) and incremental linking. It also supports adaptive sequencing, adaptation of unit sequencing, and teaching strategy. Finally, ACE implements a pedagogical agent that can give individualized recommendations to students depending on their knowledge, interests, and media preferences.
- **InterBook:** (Brusilovsky & Eklund, 1998; Brusilovsky et al., 1998): A tool for authoring and delivering adaptive electronic textbooks on the Web. InterBook supports adaptive sequencing of pages, adaptive navigation by using links annotation, and adaptive presentation. Adaptive sequencing and navigation are implemented by using a frames-based presentation that includes a partial and adaptive table of contents, a presentation of the prerequisite knowledge for the current page, and overview of the concepts, which this paper discusses. For its implementation, InterBook uses the same approach and architecture as ELM-ART II.
- **DCG:** (Brusilovsky & Vassileva, 2003; Vassileva & Deters 1998): An authoring tool for adaptive courses. It generates personalized courses according to the student's goal and model, and dynamically adapts the course content according to the acquired knowledge. DCG supports adaptive sequencing by using a domain concept structure, which helps in generating a plan of the course. DCG uses the concept structure as a roadmap for generating the course plan. A planner is used to build the course plan by searching for subgraphs that connect the concepts known by the learner to the new goal concept. The course sequencing is elaborated by linearizing the subgraphs by using the pedagogical model. The pedagogical model contains a representation of the instructional

tasks and methods and a set of teaching rules. DCG uses two instances of the student's model, one on the server side updated only after closing the learning sessions, and a more dynamic one on the client (learner) side. The learner's model is represented as an overlay with the concepts structure and contains the probabilistic estimations of the student's level of knowing the different concepts.

- **AHA:** (De Bra et al., 2002; De Bra et al., 2003): A generic system for adaptive hypermedia whose aim is to bring adaptivity to all kinds of Web-based applications. AHA supports adaptive navigation (annotation + hiding) and adaptive presentation. AHA's general structure is similar to that of the other systems discussed previously. AHA's adaptive engine consists of three parts: a domain model, a user model, and an adaptation model. The domain model describes the teaching domain in terms of concepts, pages and information fragments. It also contains the concepts' relationships. AHA uses three types of concept relationships: (1) link relationships, which represent the hypertext links among the page's concepts; (2) generated relationships, which specify the updates to the user model related to the page's access; and (3) requirement relationships, which define the prerequisites for the page's concepts. The user model consists mainly of a table presenting for each page or concept an attribute value that represents how the user relates to this concept. The AHA user model differs from other systems in that the concepts' attributes can be non-persistent and have negative values. The adaptation model consists of a collection of rules that define the adaptive behavior of AHA. Generated rules (corresponding to the generated relationships) and requirement rules (corresponding to the requirement relationships) are part of this model.
- **ILESA:** (López et al., 1998): An intelligent learning environment for the Simplex Algorithm. It implements adaptive sequencing (i.e., lesson, problem) and provides problem-solving support. ILESA follows the traditional model of an intelligent tutoring system with six components: engine, expertise module, student diagnosis module, student interface, instructional module, and problem generator. The expertise module in ILESA is a linear programming problem solver for the Simplex algorithm. The system provides a great number of different ways to solve a problem, since this system needs to allow for the diagnosis of a student's answers. The student diagnosis module provides a graph of learned skills. The domain is broken down into a list of skills for solving a Simplex problem, and a graph representing the relationships among skills is presented. The student model consists of an array of numbers representing the student's score for each of the basic skills. The problem generator is used to generate an unlimited number of problems and to provide the student with the appropriate type and level of problem. The instructional module controls the pedagogic functioning (problem posed, help offered) of the system, and coordinates the actions of the expert system, the student diagnosis module, and the problem generator. The engine contains the control mechanism that guides the behavior of the system.

FUTURE TRENDS: PERSONALIZED M-LEARNING AND ADAPTATION OF THIRD-PARTY CONTENT

In the literature, m-learning has been defined from different views. Some definitions take technology as the starting point (Farooq, 2002); other definitions (Nyiri, 2002) relate it more to distance education by focusing on the principle of anytime, anywhere, and any device. Leung (2003) identifies four characteristics for m-learning: dynamic by providing up-to-date material and resources, operating in real time by removing all constraints on time and place, adaptive by personalizing the learning activities according to the learner background, and collaborative by supporting peer-to-peer learning. M-learning is still in its birth stage, and most of the research projects are focusing on the connectivity problem of using wireless networks or the problem of accessing course content using mobile terminals (e.g., PDAs such as Compaq iPaq or WAP phones) (Baek, 2002; Houser, 2002). Few of the m-learning projects have addressed the problems of adaptation of learning tasks and personalization of course content based on a student's model, learning styles, and strategy.

Taking into consideration the nature of wireless devices and network, the personalization of m-learning services requires that more intelligence should be moved to the user terminal. New technologies such as mobile agents and Web services are promising tools for implementing adaptive m-learning services.

Unlike traditional ITS systems, new e-learning and m-learning systems should be open to third-party providers. In fact, the future trend is toward the implementation of infrastructure (i.e., e-marketplaces) that support and provide collaborative e-learning services. Thus, we need to implement a process that provides user-side device independence for content (i.e., publishers or Web content). Learning object standards (Wiley, 2000), XML, ontology, and semantic Web technology are promising tools for adapting third party content. The main idea behind the adaptation process is to construct a basic generic document from the source and then to mark up that document with appropriate tags as determined by the user profile and the device profile.

A Web course's content always involves different resources (i.e., files, database, learning objects, etc.). Therefore, the adaptation process consists of creating a Java Servlet or JSP document that connects to data sources and objects, and produces an XML document. The main idea here is to use a two-stage process for building the service: model creation service and view transformation service. The first step generates an XML document (model), and the second step translates the generated model to a rendering format (HTML, WML, etc.) that will be presented to the user. Since the rendering format depends on the devices' features and the user's preferences, the user profile and device profile will be used in this process.

This two-stage process will provide more flexibility and device independence than would be possible otherwise:

- The separation of the service model from the service view will provide us with device independence and facilitate the maintenance of the content-generation process.
- With browsers, including a W3C-compliant XSLT engine, more processing will occur on the client side and reduce the work done by the server.
- The services may be distributed over several machines, if needed, to balance the overall load.

CONCLUSION

Personalization is a crucial aspect of e-learning services and must be addressed according to three dimensions:

- **User Characteristics:** Learning style, acquired knowledge, background/experience, preferences, navigation activity, user's individual traits (personality factors, cognitive factors) and so forth.
- **Interaction Parameters:** User's goals/tasks, collaborative/cooperative, user's interests, and so forth.
- **Technology Parameters:** Device features, connection type, network state, bandwidth, and so forth.

New technologies and standardization work such as Web services architecture, semantic Web, learning object, mobile intelligent agents, and ontology are prominent tools for implementing e-learning services. However, the key issue in implementation of personalization resides in moving more intelligence from the server side to the user terminal side.

REFERENCES

Baek, Y.K., Cho, H.J. & Kim, B.K. (2002). Uses of learning objects in a wireless Internet based learning system. *Proceedings of the International Conference on Computers in Education (ICCE'02)*, Auckland, New Zealand.

Brusilovsky, P. (1995). Intelligent tutoring systems for World-Wide Web. In R. Holzapfel (Eds.), *Poster Proceedings of Third International WWW Conference* (pp. 42-45). Darmstadt, Heseen, Germany.

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction, 6*(2-3), 87-129.

Brusilovsky, P. (1999). Adaptive and intelligent technologies for Web-based education. *Special Issue on Intelligent Systems and Teleteaching, Künstliche Intelligenz, 4*, 19-25.

Brusilovsky, P. (2003). Adaptive navigation support in educational hypermedia: The role of learner knowledge level and the case for meta-adaptation. *British Journal of Educational Technology, 34*(4), 487-497.

Brusilovsky, P., & Eklund, J. (1998). A study of user model based link annotation in educational hypermedia. *Journal of Universal Computer Science. 4*(4), 429-448.

Brusilovsky, P., Eklund, J., & Schwarz, E. (1998). Web-based education for all: A tool for developing adaptive courseware. *Computer Networks and ISDN Systems, 30*(1-7), 291-300.

Brusilovsky, P., & Vassileva, J. (2003). Course sequencing techniques for large-scale Web-based education. *International Journal of Continuing Engineering Education and Lifelong Learning 13*(1-2), 75-94.

Crowder, N.A. (1959). Automatic tutoring by means of intrinsic programming. In E. Galanter (Ed.), *Automatic teaching: The state of the art* (pp. 109-116). New York: Wiley.

De Bra, P., Aerts, A., Smits, D., & Stash, N. (2002). AHA! Version 2.0, more adaptation flexibility for authors. *Proceedings of the e-Learn—World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Association for the Advancement of Computing in Education*, Montral, Quebec, Canada.

De Bra, P., et al. (2003). AHA! The adaptive hypermedia architecture. *Proceedings of the ACM Hypertext Conference*, Nottingham, UK.

Farooq, U., Schafer, W., Rosson, M.B., & Carroll, J.M. (2002). M-education: Bridging the gap of mobile and desktop computing. *Proceedings of the IEEE International Workshop on Mobile and Wireless Technologies in Education (WMTE'02)*, Vaxjo, Sweden.

Gallis, H., Kasbo, J.P., & Herstad, J. (2001). The multidevice paradigm in know-mobile—Does one size fit all? *Proceedings of the 24th Information System Research Seminar in Scandinavia*, Bergen, Norway.

Houser, C., Thornton, P., & Kluge, D. (2002). Mobile learning: Cell phones and PDAs for education. *Proceedings of the International Conference on Computers in Education (ICCE'02)*, Auckland, New Zealand.

Kay, J. (2001). Learner control. *User Modeling and User-Adapted Interaction, 11*, 111-127.

Leung, C.H., & Chan, Y.Y. (2003). Mobile learning: A new paradigm in electronic learning. *Proceedings of the the 3rd IEEE International Conference on Advanced Learning Technologies (ICALT '03),* Athens, Greece.

Litchfield, B.C., Driscoll, M.P., & Dempsey. J.V. (1990). Presentation sequence and example difficulty: Their effect on concept and rule learning in computer-based instruction. *Journal of Computer-Based Instruction, 17*, 35-40.

López, J.M, Millán, E., Pérez, J.L., & Triguero, F. (1998). Design and implementation of a Web-based tutoring tool for linear programming problems. *Proceedings of the Workshop on Intelligent Tutoring Systems on the Web at ITS'98, 4th International Conference on Intelligent Tutoring Systems*, San Antonio, Texas, USA.

Nyiri, J.C. (2002). Toward a philosophy of m-learning. *Proceedings of the IEEE International Workshop on Mobile and Wireless Technologies in Education (WMTE'02)*, Vaxjo, Sweden.

Paiva, A., & Self, J. (1995). TAG—A user and learner modeling workbench. *User Modeling and User-Adapted Interaction, 4*(3), 197-228.

Specht, M. (2000). ACE adaptive courseware environment. *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems AH2000*, Trento, Italy.

Specht, M., & Reinhard, O. (1998). ACE—Adaptive courseware environment. *The New Review of Hypermedia and Multimedia, 4*(1), 141 -161.

Tennyson, R.D., & Rothen, W. (1977). Pre-task and on-task adaptive design strategies for selecting num-

bers of instances in concept acquisition. *Journal of Educational Psychology, 69*, 586-592.

Vassileva, J. (1994). A new approach to authoring of adaptive courseware for engineering domains. *Proceedings of the International Conference on Computer Assisted Learning in Science and Engineering CALISCE'94*, Paris.

Vassileva, J., & Deters, R. (1998). Dynamic courseware generation on the WWW. *British Journal of Educational Technologies, 29*(1), 5-14.

Weber, G., & Brusilovsky, P. (2001). ELM-ART: An adaptive versatile system for Web-based instruction. *International Journal of Artificial Intelligence in Education 12*(4), 351-384.

Weber, G., & Specht, M. (1997). User modeling and adaptive navigation support in WWW-based tutoring systems. *Proceedings of the Sixth International Conference on User Modeling, UM97*, Sardinia, Italy.

Wenger, E. (1987). *Artificial intelligence and tutoring systems—Computational and cognitive approaches to the communication of knowledge*. Los Altos, CA: Morgan Kaufmann .

Wiley, D.A. (2000). Connecting learning objects to instructional design theory. A definition, a metaphore, and a taxonomy. In D.A. Wiley (Ed.), *The instructional use of learning objects*. Retrieved August 20, 2003, from *http://reusability.org/read/chapters/wiley.doc*

KEY TERMS

AHS: Adaptive Hypermedia Systems focus on adaptive presentation and adaptive navigation support. AHS uses knowledge about its users and can incorporate domain knowledge to adapt various visible aspects of the system to the user.

E-Learning: E-learning always refers to the delivery of a learning, training, or education activity by electronic means. E-learning covers a wide set of applications and processes such as Web-based learning; computer-based learning; virtual classrooms; and delivery of content via satellite, CD-ROM, audio, and videotape. In the last few years, e-learning tends to be limited to a network-enabled transfer of skills and knowledge.

ITS: Intelligent Tutoring Systems are computer-based instructional systems using AI modeling and reasoning techniques for providing a personalized learning experience. ITS systems typically rely on three types of knowledge: expert model, student model, and instructor model.

Learner Model: The learner model represents static beliefs about the learner and, in some cases, is able to simulate the learner's reasoning.

Learning Object: Learning object is mainly used to refer to a digital resource that can be reused to support learning. However, the broadest definition includes any instructional components that can be reused in different learning contexts.

LMS: Learning Management Systems refers to environments whose primary focus is the management of the learning process (i.e., registration and tracking of students, content creation and delivery capability, skill assessment and development planning, organizational resource management).

LOM: Learning Object Metadata is the IEEE standard conceptual schema that specifies the set of attributes required to describe a learning object.

M-Learning: M-learning has emerged to be associated with the use of mobile devices and wireless communication in e-learning. In fact, mobility is a most interesting aspect from an educational viewpoint, which means having access to learning services independently of location, time, or space.

Teaching Strategy: The teaching strategy consists of didactic knowledge, a set of rules that controls the adaptation and sequencing of the course.

Picture Archiving and Communication System in Health Care

Carrison KS Tong
Pamela Youde Nethersole Eastern Hospital and Tseung Kwan O Hospital, Hong Kong

Eric TT Wong
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

Radiology is the branch of medicine that deals with the diagnostic and therapeutic applications of radiation. It is often used in X-rays in the diagnosis and treatment of a disease. Filmless radiology is a method of digitizing traditional films into electronic files that can be viewed and saved on a computer. This technology generates clearer and easier-to-read images, allowing the patient the chance of a faster evaluation and diagnosis. The time saved may prove to be a crucial element in the patient's treatment process. With filmless radiology, images taken from various medical sources can be manipulated to enhance resolution, increasing the clarity of the image. Images can also be transferred internally within departments and externally to other locations such as the office of the patient's doctor. This is made possible through the picture-archiving and communication system (PACS; Dreyer, Mehta, & Thrall, 2001), which electrsonically captures, transmits, displays, and saves images into digital archives for use at any given time. The PACS functions as a state-of-the-art repository for long-term archiving of digital images, and includes the backup and bandwidth to safeguard uninterrupted network availability. The objective of the picture-archiving and scommunications system is to improve the speed and quality of clinical care by streamlining radiological service and consultation. With instant access to images from virtually anywhere, hospital doctors and clinicians can improve their work processes and speed up the delivery of patient care. Besides making film a thing of the past, the likely benefits would include reduced waiting times for images and reports, and the augmented ability of clinicians since they can get patient information and act upon it much more quickly. The creation of a permanent, nondegradable archive will eliminate the loss of film and so forth. Today, the growing importance of PACS on the fight against highly infectious disease is also identified.

BACKGROUND

PACS (Huang, 2004) started with a teleradiology project sponsored by the U.S. Army in 1983. A follow-up project was the Installation Site for Digital Imaging Network and Picture Archiving and Communication System (DIN/PACS) funded by the U.S. Army and administered by the MITRE Corporation in 1985. Two university sites were selected for the implementation—the University of Washington in Seattle and Georgetown University and George Washington University Consortium in Washington, DC—with the participation of Philips Medical Systems and AT&T. The U.S. National Cancer Institute funded one of UCLA's first PACS-related research projects in 1985 under the title Multiple Viewing Stations for Diagnostic Radiology.

The early installations of PACS in public healthcare institutions were in Baltimore Veterans Administration Medical Center (United States), Hammersmith Hospital (United Kingdom), and Samsung Medical Center (Korea). In Hong Kong, there was no PACS-related project until the establishment of Tseung Kwan O Hospital (TKOH) in 1998. The TKOH was a newly built 600-bed acute hospital with a hospital PACS installed for the provision of filmless radiological service. The design and management of the PACS for patient care will be discussed in this article. The TKOH was opened in 1999 with PACS installed. At the beginning, due to immature PACS technologies, the radiology service was operating with film printing. A major upgrade was done in 2003 for the implementation of server clustering, network resilience, liquid crystal display (LCD),

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

smart card, and storage-area-network (SAN) technologies. This upgrade has greatly improved the reliability of the system. Since November 2003, TKOH has started filmless radiology service for the whole hospital. It has become one of the first filmless hospitals in the Greater China area (Seto, Tsang, Yung, Ching, Ng, & Ho, 2003; Tsou, Goh, Kaw, & Chee, 2003).

MAIN FOCUS OF THE ARTICLE

It certainly goes without saying that most equipment is designed for reliability, but breakdowns can still occur, especially when equipment is used in a demanding environment. A typical situation is what could be called a "single-point failure." That is, the entire system fails if only one piece of equipment such as a network switch fails. If some of the processes that the system supports are critical or the cost of a system stop is too high, then building redundancy into the system is the way to overcome this problem. There are many different approaches, each of which uses a different kind of device, for providing a system with redundancy.

The continuous operation of a PACS in a filmless hospital for patient care is a critical task. The design of a PACS for such a system should be high speed, reliable, and user friendly (Siegel & Kolodner, 2001). The main frame of the design is avoiding the occurrence of any single point of failure in the system. This design includes many technical features. The technical features of the PACS installed in a local hospital include the archiving of various types of images, clustering of Web servers installed, redundancy provision for image distribution channels, and adoption of bar-code and smart-card systems. All these features are required to be integrated for effective system performance and they are described below.

ARCHIVING OF MULTIPLE IMAGE TYPES

In order to make connections with different imaging modalities, a common international standard is important. The Digital Imaging and Communications in Medicine (DICOM) standard developed by the American College of Radiology (ACR) and the National Electrical Manufacturers' Association (NEMA) is the most common standard used today. The DICOM standard is extremely comprehensive and adaptable. It covers the specification image format, a point-to-point connection, network requirements, and the handling of information on networks. The adoption of DICOM by other specialties that generate images (e.g., pathology, endoscopy, dentistry) is also planned.

The fact that many of the medical imaging-equipment manufacturers are global corporations has sparked considerable international interest in DICOM. The European standards organization, the Comitâ Europâen de Normalisation, uses DICOM as the basis for the fully compatible MEDICOM standard. In Japan, the Japanese Industry Association of Radiation Apparatus and the Medical Information Systems Development Center have adopted the portions of DICOM that pertain to the exchange of images on removable media and are considering DICOM for future versions of the Medical Image Processing Standard. The DICOM standard is now being maintained and extended by an international, multispecialty committee. Today, the DICOM standard has become a predominant standard for the communication of medical imaging devices.

WEB TECHNOLOGY

The World Wide Web (WWW) began in March 1989 at CERN (CERN was originally named after its founding body, the Conseil Europeen pour la Recherche Nucleaire, that is now called the European Laboratory for Particle Physics.). CERN is a meeting place for physicists from all over the world who collaborate on complex physics, engineering, and information-handling projects. Thus, the need for the WWW system arose from the geographical dispersion of large collaborations and the fast turnover of fellows, students, and visiting scientists who had to get up to speed on projects and leave a lasting contribution before leaving.

Set off in 1989, the WWW quickly gained great popularity among Internet users. For instance, at 11:22 a.m. of April 12, 1995, the WWW server at the SEAS (School of Engineering & Applied Science) of the University of Pennsylvania responded to 128

requests in 1 minute. Between 10:00 and 11:00, it responded to 5,086 requests in 1 hour, or about 84 requests per minute. Even years after its creation, the Web is constantly maturing: In December 1994 the WWW was growing at roughly 1% a day—a doubling in a period of less than 10 weeks (Berners-Lee, 2000).

The system requirements for running a WWW server (Menasce & Almeida, 2001, 2004) are minimal, so even administrators with limited funds had a chance to become information providers. Because of the intuitive nature of hypertext, many inexperienced computer users were able to connect to the network. Furthermore, the simplicity of the hypertext markup language, used for creating interactive documents, has allowed many users to contribute to the expanding database of documents on the Web. Also, the nature of the World Wide Web provided a way to interconnect computers running different operating systems, and display information created in a variety of existing media formats. In short, the Web technology provides a reliable platform for the distribution of various kinds of information including medical images.

Another advantage of Web technology is its low demand on the Web client. Any computer running on a common platform such as Windows or Mac can access the Web server for image viewing just using Internet Explorer or Netscape. Any clinical user can carry out his or her duty anytime and anywhere within a hospital.

CLUSTERING OF DICOM WEB SERVERS

The advantage of clustering computers for high availability (Piedad & Hawkings, 2000) is that if one of the computers fails, another computer in the cluster can then assume the workload of the failed computer at a prespecified time interval. Users of the system see no interruption of access. The advantages of clustering DICOM Web servers for scalability include increased application performance and the support of a greater number of users for image distribution.

There is a myth that to provide high availability (Marcus & Stern, 2003), all that is required is to cluster one or more computer-hardware solutions. To date, no hardware-only solution has been able to deliver trouble-free answers. Providing trouble-free solutions requires extensive and complex software to be written to cope with the myriad of failure modes that are possible with two or more sets of hardware.

Clustering can be implemented at different levels of the system, including hardware, operating systems, middleware, systems management, and applications. The more layers that incorporate clustering technology, the more complex the whole system is to manage. To implement a successful clustering solution, specialists in all the technologies (i.e., hardware, networking, and software) are required. The authors used the clustering of Web servers by connecting all of the Web servers using a load-balancing switch. This method has the advantage of a low server overhead and requires no computer-processor power.

RAID TECHNOLOGY

Patterson, Gibson, and Katz (1988) at the University of California, Berkeley, published a paper entitled "A Case for Redundant Arrays of Inexpensive Disks (RAID)." This paper described various types of disk arrays, referred to by the acronym RAID. The basic idea of RAID was to combine multiple small, inexpensive disk drives into an array of disk drives, which yields performance exceeding that of a single large, expensive drive (SLED). Additionally, this array of drives appears to the computer as a single logical storage unit or drive.

The mean time between failure (MTBF) of the array will be equal to the MTBF of an individual drive divided by the number of drives in the array. Because of this, the MTBF of an array of drives would be too low for many application requirements. However, disk arrays can be made fault tolerant by redundantly storing information in various ways.

Five types of array architectures, RAID-1 through RAID-5, were defined by the Berkeley paper, each providing disk fault tolerance and each offering different trade-offs in features and performance. In addition to these five redundant array architectures, it has become popular to refer to a nonredundant array of disk drives as a RAID-0 array.

In PACS, RAID technology can provide protection for the availability of the data in the server. In

RAID level 5, no data is lost even during the failure of a single hard disk within a RAID group. This is essential for a patient-care information system. Extra protection can be obtained by using spare global hard disks for automatic protection of data during the malfunctioning of more than one hard disk. Today, most SANs for high capacity storage are built on RAID technology.

STORAGE AREA NETWORK

A storage area network (Marcus & Stern, 2003; Toigo & Toigo, 2003) is a high-speed, special-purpose network (or subnetwork) that interconnects different kinds of data-storage devices with associated data servers on behalf of a larger network of users. Typically, a storage-area network is part of the overall network of computing resources for an enterprise. A storage-area network is usually clustered in close proximity to other computing resources such as SUN (SUN Microsystems) servers, but it may also extend to remote locations for backup and archival storage using wide-area-network carrier technologies such as ATM (Asynchronous Transfer Mode) or Ethernet.

Storage-area networks use fiber channels (FCs) for connecting computers to shared storage devices and for interconnecting storage controllers and drives. Fiber channel is a technology for transmitting data between computer devices at data rates of up to 1 or 2 Gbps and 10 Gbps in the near future. Since fiber channel is 3 times as fast, it has begun to replace the small computer system interface (SCSI) as the transmission interface between servers and clustered storage devices. Another advantage of fiber channel is its high flexibility; devices can be as far as 10 km apart if optical fiber is used as the physical medium. Standards for fiber channel are specified by the Fiber Channel Physical and Signaling standard, and the ANSI (The American National Standards Institute) X3.230-1994, which is also ISO (International Organization for Standardization) 14165-1.

Other advanced features of a SAN are its support of disk mirroring, backup, and restoring; archival and retrieval of archived data; data migration from one storage device to another; and the sharing of data among different servers in a network. SANs can also incorporate subnetworks with network-attached storage (NAS) systems.

REDUNDANT NETWORK FOR IMAGE DISTRIBUTION

Nevertheless, all of the PACS devices still need to be connected to the network, so to maximize system reliability, a PACS network should be built with redundancy (Jones, 2000). To build up a redundant network (Marcus & Stern, 2003), two parallel gigabit-optical fibers were connected between the PACS and the hospital networks as two network segments using four Ethernet switches. The Ethernet switches were configured in such a way that one of the network segments was in active mode while the other was in standby mode. If the active network segment fails, the standby network segment will become active within less than 300 ms to allow the system to keep running continuously.

BAR-CODE SYSTEM

Recognizing that manual data collection and keyed data entry are inefficient and error prone, bar codes evolved to replace human intervention. Bar codes are simply a method of retaining data in a format or medium that is conducive to electronic data entry. In other words, it is much easier to teach a computer to recognize simple patterns of lines, spaces, and squares than it is to teach it to understand written characters or the English language. Bar codes not only improve the accuracy of entered data, but also increase the rate at which data can be entered.

A bar-code system includes printing and reading the bar-code labels. In most hospital information systems, the bar-code system has commonly been adopted as a part of the information system for accurate and fast patient-data retrieval. In PACS, bar-code labels are mostly used for patient identification and DICOM accession. They are used to retrieve records on patient examinations and studies.

SMART-CARD SYSTEM

A smart card is a card that is embedded with either a microprocessor and a memory chip or only a memory chip with nonprogrammable logic. The microprocessor card can add, delete, and otherwise manipulate information on the card, while a memory-chip card, such as

prepaid phone cards, can only undertake a predefined operation. Smart cards, unlike magnetic-stripe cards, can carry all necessary functions and information on the card. Smart cards can also be classified as contact and contactless types. The contactless smart card communicates with the reader using the radio frequency (RF) method.

In PACS, a contactless smart-card system was installed for the authentication of the user. The information about the user name, log-in time, and location are stored in a remote server through a computer network.

NO-FILM POLICY

No film was printed when the patients were still under hospital care. Film was printed only when the patient was transferred to another hospital. Under the no-film policy, the chance of spreading infectious diseases through film is reduced.

EMBEDDED LCD MONITOR

To display medical images in the hospital, LCD monitors were installed on the walls in ward areas adjacent to existing light boxes. LCD displays utilize two sheets of polarizing material with a liquid crystal solution between them. An electric current passed through the liquid causes the crystals to align so that light cannot pass through them. Each crystal, therefore, is like a shutter, either allowing light to pass through or blocking the light. Monochrome LCD images usually appear as blue or dark-grey images on top of a greyish-white background. Colour LCD displays use two basic techniques for producing colour: Passive matrix is the less expensive of the two technologies. The other technology, called thin film transistor (TFT) or active matrix, produces colour images that are as sharp as traditional CRT (Cathode Ray Tube) displays, but the technology is expensive. Recent passive-matrix displays using new colour super-twist nematic (CSTN) and double-layer super-twisted nematic (DSTN) technologies produce sharp colours rivaling active-matrix displays.

Most LCD screens used are transmissive to make them easier to read. These are a type of LCD screens in which the pixels are illuminated from behind the monitor screen. Transmissive LCDs are commonly used because they offer high contrast and deep colours, and are well suited for indoor environments and low-light circumstances. However, transmissive LCDs are at a disadvantage in very bright light, such as outdoors in full sunlight, as the screen can be hard to read. In PACS, the LCD monitors were installed in pairs for the comparison of a large number of medical images. They were also configured in portrait mode for the display of chest X-ray CR (computed radiography) images.

IMPLEMENTATION

In the design of the TKOH PACS (Figure 1), all computed tomographic (CT), magnetic resonance (MR), ultrasound (US) and computed radiographic images were archived in image servers of the PACS (Figure 2). During the diagnosis and monitoring of patients with highly infectious diseases, CT and CR scans were commonly used for comparison. A large storage capacity for the present and previous studies was required. The capacity of the image servers designed was about 5 terabytes using 2.3-terabyte SAN technology and a DICOM compression of 2.5. The image distribution to the clinicians was through a cluster of Web servers, which provided high availability of the service. The connection between the PACS and the hospital network was through a cluster of automatic fail-over switches as shown in Figure 3. Our users can use a Web browser for X-ray-image viewing for the diagnosis or follow-up of patients. The Web-based X-ray-image viewers were set up on the computers in all wards, intensive care units, and specialist and outpatient departments to provide a filmless radiological service. The design of the computers for X-ray-image viewing in wards is shown in Figure 4. These computers were built using all the above technologies for performance and reliability.

After 10 months of filmless radiological operation in TKOH, less than 1% of the cases required special X-ray film for follow-up. Basically, X-ray-image viewing through a computer network was sufficient for the radiological diagnosis and monitoring of patients. Furthermore, filmless radiology (Siegel & Kolodner, 2001) service definitely reduced the chance for spreading highly infectious diseases through health-care staff. No staff member from the radiology department became infected during the outbreak of the severe acute respiratory syndrome (SARS) in 2003. No film-loss and film-waiting times were recorded.

Figure 1. X-ray imaging modalities in the TKOH PACS

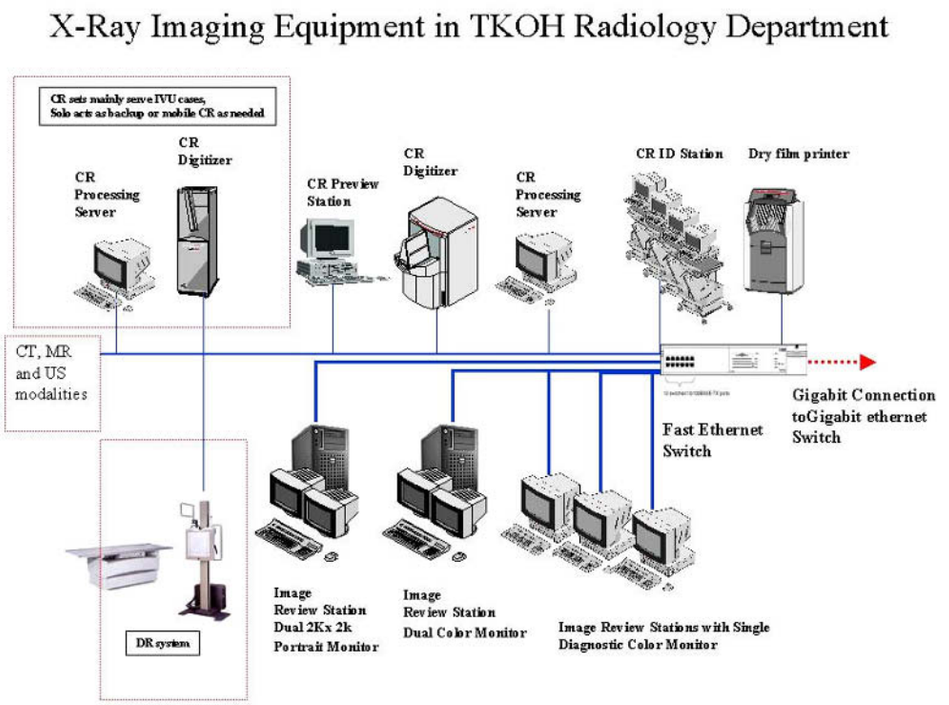

X-Ray Imaging Equipment in TKOH Radiology Department
CR sets mainly serve IVU cases,
Solo acts as backup or mobile CR as needed
CR Digitizer
CR Processing Server
CR Preview Station
CR Digitizer
CR Processing Server
CR ID Station
Dry film printer
CT, MR and US modalities
Gigabit Connection toGigabit ethernet Switch
Fast Ethernet Switch
DR system
Image Review Station Dual 2Kx 2k Portrait Monitor
Image Review Station Dual Color Monitor
Image Review Stations with Single Diagnostic Color Monitor

Figure 2. Design of the TKOH PACS

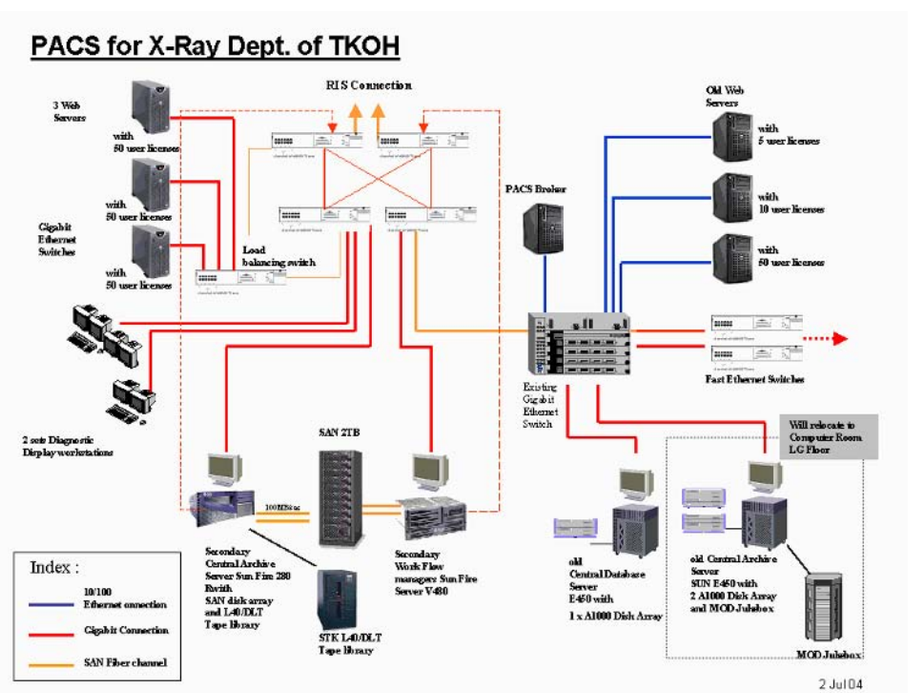

PACS for X-Ray Dept. of TKOH
RIS Connection
3 Web Servers
with 50 user licenses
with 50 user licenses
with 50 user licenses
Gigabit Ethernet Switches
Load balancing switch
2 sets Diagnostic Display workstations
SAN 2TB
100MBit/sec
Secondary Central Archive Server Sun Fire 280 R with SAN disk array and L40/DLT Tape library
STK L40/DLT Tape library
Secondary Work Flow managere Sun Fire Server V-480
Old Web Servers
with 5 user licenses
with 10 user licenses
with 50 user licenses
PACS Broker
Existing Gigabit Ethernet Switch
Fast Ethernet Switches
Will relocate to Computer Room LG Floor
old Central Database Server E450 with 1 x A1000 Disk Array
old Central Archive Server SUN E450 with 2 A1000 Disk Array and MOD Jukebox
MOD Jukebox
Index :
10/100 Ethernet connection
Gigabit Connection
SAN Fiber channel
2 Jul 04

Figure 3. Design of a PACS and hospital network interface

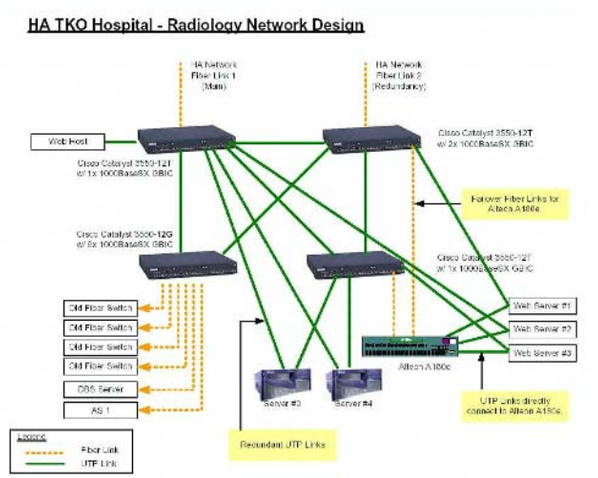
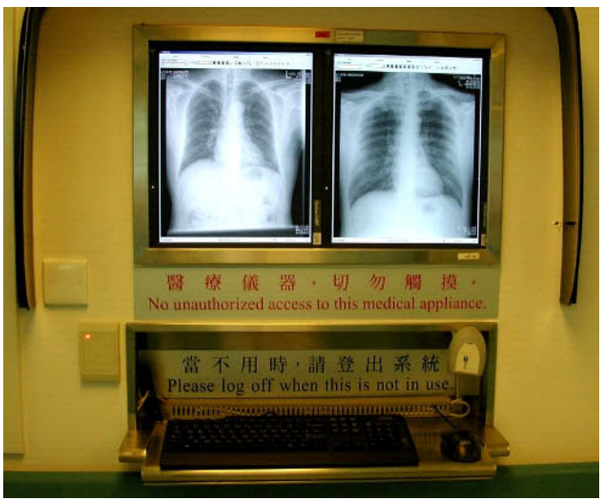


Figure 4. X-ray image viewer in wards



FUTURE TRENDS

In PACS, most of the hard disks used in the RAID are expensive fiber-channel drives. Some RAID manufacturers are designing their RAID controllers using mixed ATA (Advanced Technology Attachment) and fiber-channel drives in the same array with 100% software compatibility. This design has many benefits. It can reduce the data backup and restore from seconds to hours, keep more information online, reduce the cost of the RAID, and replace the unreliable tap devices in the future. Another advanced development of PACS was in the application of voice recognition (Dreyer et al., 2001) in radiology reporting, in which the computer system was able to automatically and instantly convert the radiologist's verbal input into a textual diagnostic report. Hence, the efficiency of diagnostic radiologists can be further improved.

CONCLUSION

It has been reported (Siegel & Kolodner, 2001) that filmless radiological service using PACS could be an effective means to improve the efficiency and quality of patient care. Other advantages of filmless radiological service are infection protection for health-care staff and the reduction of the spreading of disease through the distribution of films. In order to achieve the above tasks, many computer and multimedia technologies such as the Web, SAN, RAID, high availability, LCD, bar code, smart card, and voice recognition were applied. In conclusion, the applications of computer and multimedia technologies in medicine for efficient and quality health care is one of the important areas of future IT development. There is no boundary and limitation in this application. We shall see doctors learning and using computers in their offices and IT professionals developing new medical applications for health care. The only limitation we have is our imagination.

REFERENCES

Berners-Lee, T. (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web*. San Francisco, CA: HarperBusiness.

Dreyer, K. J., Mehta, A., & Thrall, J. H. (2001). *PACS: A guide to the digital revolution* (1st ed.). New York: Springer-Verlag.

Huang, H. K. (2004). *PACS and imaging informatics: Basic principles and applications* (2nd ed.). Hoboken, NJ: Wiley-Liss.

Jones, V. C. (2000). *High availability networking with Cisco* (1st ed.). Boston: Addison Wesley Longman.

Marcus, E., & Stern, H. (2003). *Blueprints for high availability* (2nd ed.). Indianapolis: Wiley.

Menasce, D. A., & Almeida, V. A. F. (2001). *Capacity planning for Web services: Metrics, models, and methods* (2nd ed., chap. 5). Upper Saddle River, NJ: Prentice Hall PTR.

Menasce, D. A., & Almeida, V. A. F. (2004). *Performance by design: Computer capacity planning by example* (chap. 6). Upper Saddle River, NJ: Pearson Education.

Patterson, D., Gibson, G., & Katz, R. H. (1988). A case for redundant arrays of inexpensive disks (RAID). *ACM SIGMOD Record, 17*(3), 109-116.

Piedad, F., & Hawkings, M. (2000). *High availability: Design, techniques and processes* (chap. 8). Upper Saddle River, NJ: Prentice Hall PTR.

Seto, W. H., Tsang, D., Yung, R. W., Ching, T. Y., Ng, T. K., & Ho, M. (2003). Effectiveness of precautions against droplets and contact in prevention of nosocomial transmission of severe acute respiratory syndrome (SARS). *Lancet, 361*, 1519-1520.

Siegel, E. L, & Kolodner, R. M. (2001). *Filmless radiology* (chap. 5). New York: Springer-Verlag.

Toigo, J. W., & Toigo, M. R. (2003). *The holy grail of network storage management* (1st ed., chap. 3). Upper Saddle River, NJ: Prentice Hall PTR.

Tsou, I. Y. Y., Goh, J. S. K., Kaw, G. J. L., & Chee, T. S. G. (2003). Severe acute respiratory syndrome: Management and reconfiguration of a radiology department in an infectious disease situation. *FRCR* Radiology*, 229*, 21-26.

KEY TERMS

Clustering: A cluster is two or more interconnected computers that create a solution to provide higher availability, higher scalability, or both.

Computed Radiography (CR): Computed radiography is a method of capturing and converting radiographic images into a digital form. The medium for capturing the X-ray radiation passing through the patient and generated by a standard X-ray system is a phosphor plate that is placed in a standard-size cassette, replacing the regular radiographic film. The X-ray exposure forms a latent image on a phosphor plate that is then scanned (read or developed) using a laser-beam CR reader. The CR unit displays the resultant digital image on a computer-monitor screen. By the end of the short process, the phosphor plate is erased and ready for another X-ray image exposure.

Computed Tomography (CT): Computed tomography is a specialized radiology procedure that helps doctors see inside the body. CT uses X-rays and computers to create an image. The images show up as a cross-sectional image.

Digital Imaging and Communications in Medicine (DICOM): Digital Imaging and Communications in Medicine is a medical image standard developed by the American College of Radiology and the National Electrical Manufacturers' Association.

Picture-Archiving and Communication System (PACS): A picture-archiving and communication system is a system used for managing, storing, and retrieving medical image data.

Redundant Arrays of Inexpensive Disks (RAID): RAID is a method of accessing multiple individual disks as if the array were one larger disk, spreading data access out over these multiple disks, thereby reducing the risk of losing all data if one drive fails and improving access time.

Severe Acute Respiratory Syndrome (SARS): Severe acute respiratory syndrome is a newly emerged infectious disease with moderately high transmissibility that is caused by a coronavirus.

Storage-Area Network (SAN): A storage-area network is a networked storage infrastructure (also known as a fabric) that provides the any-to-any connectivity between servers and storage devices, such as RAID disk systems and tape libraries.

Plastic Optical Fiber Applications

Spiros Louvros
COSMOTE S.A., Greece

Athanassios C. Iossifides
COSMOTE S.A., Greece

Dimitrios Karaboulas
University of Patras, Greece

Stavros A. Kotsopoulos
University of Patras, Greece

INTRODUCTION

Nowadays cabling based on symmetrical copper cables is dominant in almost all telecom applications; glass fibers predominate in long-distance networks. Whereas just a few years ago l0-Mbit/s Ethernet (l0BaseT) had the main share of interfaces in star or tree structures, today's pure star networks are predominantly set up on the basis of 100-Mbit/s connections.

Plastic optical fiber (POF) is a promising candidate for optical cabling infrastructures due to its low price, large cross-section area, easy connection and coupling with optical sources, and simple use (Daum, Krauser, Zamzow, & Ziemann, 2002).

Connecting electronic devices to the electric circuit and through data networks with copper cables always produces loops that can act as antennas or even create undesired current paths. In commercial use, these problems should always be taken into consideration. Above all, the problem of induction, for example, caused by lightning striking, has to be solved by means of appropriate protective grounding. In such a case, POF would be an interesting alternative that could surely be used in special applications. Practical and proven solutions do exist for copper cables, too.

This article is developed in four sections. In the first section, POF technical details are exposed in order to introduce the reader to the main differences among the most popular glass fibers. In the next section, several standards and bodies are described. The reader should be aware about the standards and what specifications are included. Then, applications are exposed, and finally several POF research clubs all over the world are mentioned.

POF TECHNICAL BACKGROUND

POF is a promising optical fiber and in certain applications is superior to the most popular glass optical fibers. The advantages of POF are the following:

- **Large fiber cross-section area:** The core to cladding ratio is 980:1,000 ¼m. Due to the large fiber cross section, the positioning of POF at the transmitter or receiver presents no great technical problem, in contrast to glass optical fiber.
- **Relative immunity to dust:** Particularly in industrial environments, where dust is a main problem during construction, the large fiber diameter proves to be an advantage. If dust gets into the fiber end face, it affects the input and output optical power in every case. But with POF, minor contamination does not necessarily result in failure of the transmission route. For this reason, POF can readily be connected on site in an industrial environment.
- **Simple use (great resistance to mechanical damage):** The 1-mm thick optical fiber is easier to handle, resulting in less problematic handling during installation and applications. Bending is not a serious problem and flexibility is increased in contrast to glass fibers where bending tends to break glass and attenuation is considerably increased.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

- **Low cost:** According to previous statements, the components for connection to transmitters and receivers are relatively economical. The uncomplicated processing of the end phases can be performed in an extremely cost-effective way, especially after assembling in the field.

There are, however, certain disadvantages, considering the most common applications of optical fibers.

- **Optical attenuation**: The attenuation of plastic components consisting of POF is extremely large, resulting in short-distance applications in telecommunications and industry (Daum, Brockmayer, & Goehlich, 1993).
- **Low supported data rate:** Due to the large core cross-section area, a lot of modes are supported during transmission, resulting in a considerable time dispersion. As a result, the data rate is considerably reduced (Gunther, Czepluch, Mader, & Zedler, 2000).
- **Low bandwidth-distance product:** Considerable data rates for telecommunications and industrial applications are achieved for short-distance connections (<500 m).

The plastic materials used in POF are polycarbonate core material (PC), polystyrene core material (PS), polymethylmetacrylate core material (PMMA), and fluropolymers for cladding materials. These materials have different optical windows for low-attenuation applications, according to Table 1.

STANDARDS

ATM Forum

In several documents of the ATM Forum (asynchronous transfer mode; ATM96a, ATM96b, ATM97, ATM99), the transmission medium for data transmission with 155 Mbit/s up to 50 m with POF or 100 m with HPCF (hard plastic-clad fiber) respectively is described. According to the last document from January 1999 (AF-PHY-0079.001, ATM99), the attenuation of a connection with POF should not be greater than 17 dB, of which 4 dB represents the connector loss. With HPCF connections, the maximum attenuation amounts to 6.5 dB, of which 4.5 dB contains the plug attenuation. The polymer optical fiber with a diameter of 1,000 ¼m has a step-index profile as specified in IEC 61793-2 Section 4 cat A4d. The HPCF is a 225-¼m multimode, step-index hard polymer-clad fiber as specified in IEC 61793-2 Section 3 cat A3D. The minimum bandwidth, because of mode dispersion, amounts to 10 MHz/km measured at 650 nm in accordance with 61793-1-C2A or IEC 61793-1-C2B.

.

IEEE 1394b

In the document P1394b, the characteristics of POF and HPCF cables for data rates up to 125 Mbit/s (S100) and 250 Mbit/s (S200) at transmission wavelengths of 650 nm are specified. By using these transmission media, the goal is to provide economical point-to-point connections between IEEE 1394 components for 50 m (POF) or 100 m (HPCF) respectively.

Table 1. Optical window for different materials

Material	Refractive core index n	Optical attenuation
PMMA	1.49	70-100 dB/km at 570 nm 125-150 dB/km at 650 nm
PC	1.58	700 dB/km at 580 nm 600 dB/km at 765 nm
PS	1.59	90 dB/km at 580 nm 70 dB/km at 670 nm

SERCOS

SERCOS (serial real-time communication system) describes a standardized digital interface for data communication in industrial CNC (computer numerical control) applications. It enables a serial real-time communications system that consists of optical point-to-point connections in a ring structure. Polymer optical fibers up to a length of 60 m with LED (light emitting diodes) transmitters in the wavelength range of 640 nm to 670 nm are used with a data rate of 2 Mbit/s. Detailed information can be found at the Web site http://www.sercos.org.

Profibus

The Profibus is a field-bus system that is standardized in EN 50170 Volume 2. A field bus characterizes a type of network on the lowest level of automation directly in the technical process (DIN 19245). The maximum number of users per network segment is limited to 32. In addition to shielded two-wire cables, polymer optical fibers are also specified as a transmission medium. Depending on the transmission medium, various network topologies are permitted. Using POF, a transmission length of up to 60 m is attained. The snap-in plug system by Hewlett Packard is used for plug installation. A further addition is the introduction of an optical interface for POF with a data rate of 12 Mbit/s.

INTERBUS

The Interbus system describes a full-duplex data transmission in a ring structure. Copper cables as well as different types of fiber waveguides, for example, polymer optical fibers, HPCF fibers, and multimode glass fibers, can be employed. With POF, distances of up to 70 m can be bridged with HPCF fibers of 400 m and glass fibers up to 3,600 m long. The optical part of the Interbus system is treated in detail in the "Technical Guideline: Optical Transmission Engineering" (Int97).

D2B

The D2B (domestic digital bus) standard specifies a ring system that connects different devices in vehicles such as navigation computers, car radios, CD changers, telephones, and so forth using POF.

MOST

The MOST (media-oriented system transport) specification gives recommendations for multimedia-capable networks in automobiles of the future. Since the founding of the MOST initiative in 1998, 14 international automobile manufacturers together with 50 key component suppliers have been working on the MOST technology. Polymer optical fibers are used as the transmission medium.

APPLICATIONS OF POLYMER OPTICAL FIBERS

The polymer optical fiber has been employed for many years now. In lighting technology, it represents a widely used and recognized medium. The share of POF in sensor technology and data communications, on the other hand, is rather small. Up until now, POF was predominately used in niche areas. At present, a dramatic change is taking place, especially in data communications. The digitalization of diverse entertainment media (music, language, video, pictures), a process which is practically completed, and the steadily growing amount of terminal equipment has led to a massive demand for reasonably priced, fast, and reliable data connections in all areas of private and public life. In this article, some possible uses for POF, as well as areas in which it is already being utilized, will be discussed. Those areas of use that lie outside the field of data communications and beyond the scope of this article will only be treated briefly (Marcou, 1997).

POF in the Automotive Field

In Europe, the use of polymer optical fibers for the entertainment networks in DaimlerChrysler vehicles since 1998 represents the first comprehensive application of POF in data communications ("Recommended environmental practices," 1978). The following arguments speak for the use of POF in vehicles:

- low cable weight
- small cross section
- insensitivity to electromagnetic disturbances

In vehicles, airplanes, and rail transportation, more and more digital communications connections are being utilized. As a result, increased demands on the architecture of the data connections as well as the transmission media are being made. In the area of driver information and entertainment systems, less relevant in regard to safety requirements, serial bus systems are being increasingly used. The individual devices are switched in series by means of high-rate connections. The advantage here is the saving of cables. The disadvantage is the breakdown of an entire series of devices when a transceiver subassembly is defective.

In the meantime, vehicles are equipped with additional devices such as navigation systems (Navi), traffic guidance systems (telematics), mobile Internet access, and DVD players.

The data structure of the MOST system is oriented extensively toward the demands of the connectable multimedia terminal devices. The sampling frequency of a CD player at 44.1 kHz forms the basis of the bus clock pulse. The formation of blocks, each with 512 bits, results in a gross baud rate of 22.6 Mbit/s. Principally, all components are designed for a frame clock-pulse rate of between 30 and 50 kHz. Up to 64 nodes can be connected in a MOST link.

A differentiation is made between synchronous data (with permanently assigned channels), asynchronous data (using available channels), and control data, with permanent assignment within a time frame. Because of the frame clock-pulse rates, the delay times are at a maximum of 25 ¼m.

Although it was first conceived as a pure ring architecture, the MOST system can assume other topologies, for example, combined rings, star, and so forth, by adding system master units. Space requirements and the weight of the cable harnesses can be limited in spite of the increasing number of data connections. Even in the MOST specifications, different hybrid plug-and-socket connectors are planned so that the power supply and the data transmission can be installed at the same time.

POF in Data Communication

Data communication is of decisive importance in an ever-increasing number of technical fields. Data networks form the basis of telecommunications companies (Levin, Baran, Lavrova, Zubkov, Poisel, & Klein, 1999). Local data networks (LANs [local-area networks]) connect companies' computers and databases. Through the setting up of virtual networks (virtual private networks), both areas are growing together. Even in the private sector, digital data communication is becoming more important. Audio and video devices based on digital technologies are becoming increasingly network compatible (Gunther et al., 2000).

POF systems have already been created for all four interfaces mentioned. The ATM Forum has already specified the use of PMMA POF for 155 Mbit/s (Daum et al., 1993). Of particular interest is the inclusion of POF in the IEEE 1394 specification (Up until now, 100 Mbit/s and 200 Mbit/s over 50 m has been used; 400 Mbit/s over 100 m is in preparation). In contrast to Ethernet, this interface could gain acceptance not only with computers, but also in diverse multimedia devices such as game consoles, cameras and video cameras, televisions and DVD players, and with computer peripherals.

The IEEE 1394 standard is intentionally not fixed to a medium, but provides the user with the option of selecting his or her own cable. Therein lies great application potential, especially for POF as illustrated in the overview above.

In addition to the question of possible interfaces, the general building network market in Europe should be considered. In contrast to other countries such as Japan or the U.S.A., most people in Europe live in houses for several families.

POF in Lighting Technology

Polymer optical fibers are employed in great amounts in many areas of lighting technology. Two variants in particular are widely used. In the first case, polymer optical fibers are employed as pure light guides when the light source and the object to be illuminated are spatially separated (Kaino, 1985). Secondly, the POF itself is employed as a means of lighting, which is very decorative, especially for illuminating outlines.

The use of glass fibers for guiding light has been known and well established for a long time. In communications technologies, glass fibers are employed with attenuation under 1 dB/km. These consist, however, of highly pure quartz glass, the use of which for lighting technology would be prohibitively expensive. In such a case, glass fiber bundles made of

reasonably priced material are employed for greater flexibility. The glass fiber shows a clearly better performance from 600 nm and up (Kaino, 1989). There are advantages for POF, especially in the blue and green spectral range, which is important for color reproduction. The possible length of fiber bundles is effectively increased even if less light can be injected into a POF because of the lower temperature load. In addition to fiber bundles, thick polymer optical fibers for guiding light can also be used if no tight bending is necessary.

POF in Sensor Technology

In sensor technology, the great diameter of POF in many areas of application is of special interest. The two essential areas for making use of POF are the sole transmission of sensor data and, in particular, the use of the POF itself as a sensor. POF is very suitable as a distance or movement sensor in which the reflected light is measured (Kaino, 1986).

Table 2. Technologies for home networks

Technology	Performance	Advantages/disadvantages
Radio systems		
UMTS	2 Mbit/s over 70 m 300 kbit/s over some 100 m	No local cross-linking
Bluetooth	1 Mbit/s over 10 m	Very easy cross-linking; limited capacity
Wireless ATM	25 Mbit/s over 30 m	Supports several services; still relatively expensive
Copper cables		
PNA	Some Mbit/s	Requires existing telephone lines; subject to disturbances
Coaxial cables	Some 100 Mbit/s	Requires existing coaxial lines; relatively costly converters
Data cables	1 Gbit/s over 100 m	Thick cable (approx. 7 to 8 mm); most widely spread LAN technology
PLC	Some Mbit/s	Easy to install; susceptibility to disturbances and emissions; largely untested
Optical cables		
Glass SM fiber	Practically unlimited	Extremely expensive installation
Glass MM fiber	2.5 Gbit/s	Limited expense for installation
PMMA POF	Some 100 Mbit/s over 100 m	Still very new technology; extremely easy installation

POF in Domestic Applications

Today's apartments are mostly equipped with three different cable-based networks: the telephone network, a connection to the broadband cable network or an antenna system, and the 230 V electrical power supplies (Kaino, 1986). Each of these networks is adapted for its own specific, albeit very different, purpose.

The list of possible devices requiring cross-linking could be expanded at will. Surveillance and control systems for heat, windows, and doors have increasingly gained in importance. The author personally experienced apartments still being planned and built in 2001 without any kind of system for networking. The tenant is thus confronted with the problem of establishing data connections between devices with the lowest possible expenditure of time and money. Two possibilities for completely overcoming such a situation without installing cables is to use PowerLine technology or to set up a radio system. Both options are technically advanced and thoroughly affordable. However, the possible bit rates and the attainable quality are subject to definite limitations. Cable-based systems are preferable when transmitting high-quality moving pictures in real time or with a broadband connection of computers, for example, when working at home. Different copper cables as well as optic fibers can be considered. Table 2 summarizes some possible technologies for use in private surroundings.

Table 3. Interfaces for home networks

Interface	Bit rates	Advantages/disadvantages
ATM Forum	25 Mbit/s, 155 Mbit/s, 622 Mbit/s, 2.5 Gbit/s	Supports high-quality services and is already employed in long-distance networks; up until now, was too expensive for home use
Ethernet	10 Mbit/s, 100 Mbit/s, 1,000 Mbit/s	Used above all for IP (Internet protocol) applications; widespread and good value; dominant in LAN field; difficult with video transmission
USB	12 Mbit/s (new 480 Mbit/s)	Widespread standard for PCs (personal computers); very simple operation; requires running PC; up until now, data rates were too low
IEEE 1394	100 Mbit/s, 200 Mbit/s, 400 Mbit/s, 800 Mbit/s, up to 3.2 Gbit/s planned	Universal system for all applications (including video); multimaster network with extremely easy operation

As can be seen in Table 2, the PMMA POF lies in the midrange of performance characteristics for the various transmission media. In regard to the simplicity of installation, radio systems and PLC (power line communication), of course, cannot be surpassed. Among the cable-based systems, POF is distinguished as having the easiest cable setup and the most reasonably priced connection technology.

Besides the question of transmission media, the point of greatest interest is the interface to the consumer. A system can only gain general acceptance when terminal devices are equipped with appropriate connectors, the services desired can be supported with sufficient quality, and the components for setting up the network are available at reasonable prices. Table 3 lists some of the interesting interfaces.

POF CLUBS

The Japanese POF Consortium

Of the interest groups existing today in the field of polymer optical fibers, the Japanese POF Consortium can look back on the greatest amount of activity. It was founded in 1994 and has been led since by Professor Yasuhiro Koike, who has achieved international recognition for his numerous publications, especially on graded-index-profile polymer optical fibers. A total of approximately 70 institutes and manufacturers are represented in the Japanese POF Consortium.

HSPN and PAVNET

In the U.S.A., the IGI Company in Boston can be viewed as the most important representative of POF-interested parties. IGI regularly publishes POF news and sells different studies on developments within the telecommunications field, including polymer optical fibers. IGI organizes annual POF world events, which are intended primarily for commercial users. Of international importance are two consortiums in the U.S.A. that have been working in succession for several years on the development of polymer optical fiber systems, primarily for use in avionics.

The HSPN Consortium (High-Speed Plastic Network) was founded in 1994, the aim of which was the development of 650-nm VCSEL by Honeywell. PMMA-based GI POF was to be developed concurrently. Both products were able to be demonstrated under laboratory conditions, but have not yet been developed for series production.

At the end of the project in 1997, the successor organization PAVNET (Plastic Fiber and VCSEL [vertical cavity surface emitting laser] Network) was founded. The newest member is Lucent Technologies. The goals are the following.

PF-GI POF with less than 60 dB/km at 500 to 2,000 nm

Expansion of the temperature range to more than 125°C

Use of existing VCSEL technology at 850 and 1,300 nm

622 Mbit/s over 30 m, later to be expanded to 2,500 Mbit/s over 100 m

In contrast to the Japanese approach, Boston Optical Fiber used a Teflon-based material, but so far, the losses have still been in the vicinity of some l,000 dB/km.

The French POF Club

The French POF Club was founded as early as 1987. The first international POF conference took place in Paris in 1992 and was organized by IGI Europe. By 1994, approximately 200 members were registered in the French Plastic Optical Fibre Club. It is part of the French Optical Society (SFO) and is supported by the French Atomic Energy Commission. The background to this involvement is the idea of using scintillating polymer optical fibers for proof of elementary particles.

Participants include representatives from universities, research institutions, industry, government, and military institutions. There are 50 to 80 participants at the biannual meetings. In 1994, the FOP (French plastic OPtical fiber club) published the first comprehensive book on polymer optical fibers, which has been available in an English translation (Marcou, 1997) since 1997.

The Information Technology Society Subcommittee 5.4.1: Polymer Optical Fibers

In Germany, there has been considerable interest in POF for some time now, in particular through the

activities of the chemical industry (Hoechst, Bayer). Until 1996 there was no national interest group in the field. The creation of just such a group goes back to a meeting of various German participants at the POF conference in Paris (October 1996). After some preliminary preparations, it was decided by the subdepartment 5.4, Communication Cable Networks, of the Information Technology Society (ITG) within the Association for Electrical, Electronic & Information Technologies (VDE) to found the subcommittee (FG) 5.4.1, Polymer Optical Fibers, on December 3, 1996.

EU POF NET

In order to better coordinate European activities in the field of polymer optical fibers, a European information network (see http://eu-pofnet.org) is presently being set up. One of the first steps was the founding of the work group Polymer Optical Fibers within the framework of the FOToN project in Nuremberg on December 6, 2000.

REFERENCES

Daum, W., Brockmayer, A., & Goehlich, L. (1993). Influence of environmental stress factors on transmission loss of POF. *Proceedings of POF'1993*, Den Haag, July 28-29, 94-98.

Gunther, B., Czepluch, W., Mader, K., & Zedler, S. (2000). Multiplexer for attenuation measurements during POF durability testing. *Proceedings of POF'2000*, Boston, September 5-8, 209-213.

Int97 Interbusclub Deutschland e.V. (1997, November). *Technische Richtlinie: Optische: Ubertragungstechnik.* Version 1.0, Article no. 9318201.

Kaino, T. (1985). Influence of water absorption on plastic optical fibers. *Applied Optics, 24*(23), 4192-4195.

Kaino, T. (1986). Plastic optical fibers for near-infrared transmission. *Applied Physics Letters, 48*(12), 757-758.

Kaino, T. (1989). Polymers for optoelectronics. *Polymer Engineering and Science, 29*(17), 1200-1214.

Levin, V. M., Baran, A. M., Lavrova, Z., Zubkov, A., Poisel, H., & Klein, K. (1999). Production of multilayer optical fibers. *Proceedings of POF'1999,* China, July 14-16, (pp. 98-101).

Marcou, J. (1997). *Plastic optical fibers: Practical applications*. John Wiley & Sons, Masson.

KEY TERMS

ATM: Asynchronous transfer mode.

LAN: Local-area network.

Optical Attenuation: The attenuation of transmitted light through optical waveguides mostly due to material absorption.

Optical Window: A range of wavelengths in which attenuation has the lower value. The criterion to choose the optical source.

Potentials of Information Technology in Building Virtual Communities

Isola Ajiferuke
University of Western Ontario, Canada

Alexander Markus
University of Western Ontario, Canada

INTRODUCTION

In recent years, virtual communities have become the topic of countless books, journal articles and television shows, but what are they, and where did they come from? According to Preece, Maloney-Krichmar, and Abras (2003), the roots of virtual communities date back to as early as 1971 when e-mail first made its appearance on the Advanced Research Projects Agency Network (ARPANET), which was created by the United State's Department of Defense. This network would lead to the development of dial-up bulletin board systems (BBSs) which would allow people to use their modems to connect to remote computers and participate in the exchange of e-mail and the first discussion boards. From these beginnings a host of multi user domains (MUDs) and multi-user object oriented domains (MOOs) would spring up all over the wired world. These multi-user environments would allow people to explore an imaginary space and would allow them to interact both with the electronic environment and other users. Additionally, listservs (or mailing lists) sprang up in 1986, and now, almost two decades later, they are still in use as the major method of communication among groups of people sharing common personal or professional interests (L-Soft, 2003). Since then the Internet has exploded due to the development of Web browsers as well as the development of communications technologies such as broadband, digital subscriber line (DSL), and satellite communications. Groups of people from as few as two and reaching to many thousands now communicate via email, chat, and online communities such as the Whole Earth 'Lectronic Link (WELL) and such services as MSN, Friendster, America Online (AoL), Geocities, and Yahoo! Groups. Other examples of online communities are collaborative encyclopedias like Wikipedia. Web logs (Blogs) like Slashdot.com and LiveJournal allow users to create their own content and also to comment on the content of others. They also allow the users to create identities and to make virtual "friends" with other users. The definition of virtual community itself becomes as convoluted as the multitude of technologies that drives it. Are e-mail lists, message boards, and chat rooms online communities or are they virtual communities? Virtual communities might be persistent worlds as those found in popular online games (Everquest, 2004, Ultima Online, 2004) or virtual worlds (such as MUDs and MOOs) where the user is able to explore a simulated world or to take on a digital "physicality" in the form of an avatar. It becomes clear from the literature that the terms are still used interchangeably.

BACKGROUND

In the online version of his 1993 book on virtual community, Rheingold states that:

When you think of a title for a book, you are forced to think of something short and evocative, like, well, 'The Virtual Community', even though a more accurate title might be: 'People who use computers to communicate, form friendships that sometimes form the basis of communities, but you have to be careful to not mistake the tool for the task and think that just writing words on a screen is the same thing as real community.(http://www.rheingold.com/vc/book/intro.html)

And it is with this statement about his own work, that Rheingold so eloquently captures the essence of the problem in defining the virtual community and

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

separating it from the technology. That is to say, there is a difference between the online community and what is known as computer mediated communication (CMC). According to December (1996), CMC deals with both a technological and a sociological approach to communication over the Internet, but it is important to note that because a communication happens "online", it does not imply community. Etzioni and Etzioni (1999, pp. 241) touch on this issue precisely when they say "that to form and sustain communities, certain conditions must be met". They state that in the CMC literature, community can refer to either tightly knit social groupings or it can also mean people using a common service whom have no ties to each other (such as subscribers to a news source).

The difficulty is in defining the term community itself. Both the Mirriam-Webster's online dictionary (2004) and the Wikipedia online (2004) define community as having, at its heart, a common bond between the members, originally having some geographic significance (such as the same city or even the same nation) and more recently to include more widely spaced persons who form either a physical or a digital Diaspora. Those that have attempted to define or explore aspects of virtual community include Rheingold (2000), Preece (2000), Fernback and Thompson (1995), and Wilson and Peterson (2002). However, there are also those that do not share an optimistic view of virtual communities with some seeing CMC as necessarily precluding community without face to face communication (Weinreich, 1997, Shenk, 1998). The idea of virtual community strikes at the very heart of what we know a community is composed of and the presence of a wide range of views indicates that we will not have a definitive answer for some time to come. It is possible that in the future, those who have grown up with the Internet and have been "wired" from an early age might be more willing to accept the virtual community in the same way that we accept our own physical communities. We don't even think about it, we just go on about our lives participating in the communities that we align ourselves with. In 20 years will there even be a need to define virtual or online communities at all?

MAIN THRUST OF THE ARTICLE

The day of a typical graduate student often starts with checking his e-mail, sometimes even before taking a shower or drinking a cup of coffee. He is plugged into the World Wide Web, and there's no need to turn on his computer because it's always on. While still thinking about how best to start his day, he is reading news from the Cable News Network (CNN), and answering any number of emails from friends, family, and students. These aren't really communal actions; they don't constitute a community in any sense. Arguably, his e-mail activity allows him to communicate with people in his physical community, but it is an extension of the community in which he lives; it is online, but it is not virtual.

In the long list of e-mails he has received might be some postings to his professional listserv. Many of the people posting to this list are in his professional community, he knows them by name, a number of them he knows in person, he sees them at conferences, they have come to give guest lectures at his university, and some of them are his colleagues. When he replies to these messages, he is still extending his physical community, but there is also a sense of virtuality included in the exchange because some of these people he knows only in the online context, yet he feels he actually knows them.

Some virtual communities are also served by online message boards, such as the message board for people who suffer from depression. The board is full of discussions about possible side effects, battling depression in general, common advice, and support between members of the message board. This is where the virtual community begins to take shape. Hundreds of people separated by geography, race, religion, or creed, but bound by their common battle against depression. Some of them have been posting only for a few days, others since 1995 (Wing of Madness, 2004). It is this persistent online location in the form of the message board that allows for the creation of the virtual community. As in every community people come and go as time marches on and they all participate to varying degrees ranging from active participation to passive lurking (Nonnecke and Preece, 2003). Lurkers, as they are called, can comprise as much as 99 percent of an online community though they receive the communications from

the community, they offer none in return and are, for all intents and purposes, invisible.

Virtual communities are of particular interest to researchers in the establishment of communities of practice and in health care applications. Hara and Kling (2002) and Millen and Dray (2000) illustrate the need for online communities in the process of knowledge creation and the importance of virtual communities to professionals. Some of the applications could include medical researchers such as those interested in training people to diagnose diseases (Utzinger, Tanner, & Singer, 2001). Utzinger and his colleagues seek to halve the current malaria burden world wide by the year 2010. It is their belief that this can be accomplished by establishing an online community that could be used to educate and train people to diagnose this condition, with the end result being that both treatment and diagnosis are greatly improved. The advantage of this particular method lies in the leveraging of CMC to facilitate the sharing of information and for the promotion of sharing between dispersed practitioners. Similarly, Kodama (2001) believes that video conferencing can be used to supplement and to make accessible health care from home. Recent advances in broadband Internet connections have made this possible.

Virtual communities have also challenged notions of democracy and politics. According to Shenk (1998, p. 53):

People just don't understand how tumultuous this technological revolution is going to be. They think the world will look pretty much the way it does now, just faster. But they don't get it: It's going to be a completely different world. I'd say democracy has about a fifty-fifty chance of survival.

Shenk argues further that democracy can be stifled through online communities; his argument is that since those communities are insular, and because they tend to attract people with like minded views, that the scope of shared discussion is diminished as a result. In the public forum, people with dissenting views would possibly speak up in light of disturbing proclamations, which would in turn, generate free and open discussion. Shenk believes that online communities lack this voice of dissention. This argument runs counter to Rheingold's (http://www.rheingold.com/vc/book/3.html) assertion that:

The ability of groups of citizens to debate political issues is amplified enormously by instant, widespread access to facts that could support or refute assertions made in those debates. This kind of citizen-to-citizen discussion, backed up by facts available to all, could grow into the real basis for a possible electronic democracy of the future.

The truth is that the technology can be used either way, depending on the structure of the software, the interface, and the language used. It is also subject to influence by the country in which it is located, the laws and the amount to which speech is free and to the extent that the virtual community is moderated. What can be concluded with some certainty is that CMC and virtual communities are just that, they are virtual; they are surrogates for real world communication. Not everyone has the same comfort level with these communities, and that they are not freely accessible as is evident by the digital divide.

Virtual communities also provide opportunities for distance education and learning. They allow people with disabilities, those in remote areas, and those who cannot attend traditional educational facilities for whatever reason the opportunity to learn in a virtual classroom. For the United Nations Educational, Scientific and Cultural Organization (UNESCO), the ability to use virtual classrooms also gives teachers in all countries a greater range of options for developing their skills as well (UNESCO, 2001). The American Council on Education Center for Policy Analysis also states that "virtual" schools could be called "click" universities in that non-traditional populations such as adult learners, and part-time students could benefit from such services (Levine & Sun, 2002). The use of online communities for education is inherently problematic in that they depend on the user being literate in both in the language of instruction and in the use of computers. Though the latter may not be as large an issue in North America where elementary students are being exposed to computers at an early age, the impact would be far greater in developing nations where access to computers and ultimately the Internet are not as available.

Virtual communities are almost ubiquitous in nature at the present time, and are sure to become so in the future. The more we think about them, the more we realize how prevalent they already are on

the Internet. In short, they are a part of our lives, now and in the future. It is through our understanding of how virtual communities are created, propagated and how we as individuals find our place within them that we will be better able to leverage them to our benefit in the future.

FUTURE TRENDS

As the technology that drives the Internet advances, so will the realm of CMC advance and the ability to which we can form and grow virtual communities (Wellman, 2001). As access time decreases and the bandwidth to stream voice and video over the Internet increases so will the possibilities for online communities be magnified. Technologies such as haptics may provide additional depth to existing virtual communities and may provide for additional accessibility to them (Laboratory for Intelligent Mechanic Systems, 2004). Virtual or online communities cannot simply be created or destroyed according to the whims of the programmer or the politician (Shirky, 2003). The case of a virtual community is not simply a "if we build it, they will come" phenomenon. They are constantly being created and dying out as their usefulness expires.

Little is known about how virtual communities are created and what makes them successful or not. More accurately perhaps, there is no single list of characteristics that if presented to a user will prompt them to become a member of the virtual community. This is the difference between users (people who come to a website to surf) and members (those people that take an active role in, and/or identify with an online community). Much remains to be learned about how online communities operate and how they are affected by region, politics, language, and socioeconomic factors. Little is known about the motivations of the users themselves and what drives them to participate. Perhaps the problem with setting out to create virtual communities lies in the formation of an artificial environment. Future research may shed light on these issues.

CONCLUSION

Virtual communications are over 30 years old, and since the creation of ARPANET in the 1970's, the technology has made the creation of online communities possible. Though there may be debates for some time to come about the distinctions between online and virtual communities, we can be assured that they are here to stay. They serve to unite people of all kinds, in all places, through the digital universe that is CMC. They are used for personal interest, professional development, education, health care, democracy, and as tools of resistance against oppressive regimes. Virtual communities unite people scattered across the globe in Diaspora and allow them to contact each other. Not every digital communication constitutes or creates community, but they serve to strengthen the bonds that already exist or to create new ones. In the future, increased bandwidth and integrated technologies like video satellite phones will further increase the opportunities for such communities to be created. As computer technology advances, users will find themselves more plugged in, and will increasingly represent themselves in digital form. These forms could include live video feeds of themselves, or in the case of virtual worlds, 3-dimensional avatars that represent the user in the digital environment.

REFERENCES

December, J. (1996). Units of analysis for Internet communication. *Journal of Communication, 46*(1), 14-38.

Etzioni, A. & Etzioni, O. (1999). Face-to-face and computer-mediated communities: A comparative analysis. *The Information Society, 15*, 241-248.

Everquest (2004). EverQuest: You're in our world now! Retreived March 1, 2004, from *http://everquest.station.sony.com/*

Fernback, J. & Thompson, B. (1995). Virtual communities: Abort, retry, failure? Retrieved March 1, 2004, from *http://www.well.com/user/hlr/texts/VCcivil.html*.

Hara, N. & Kling, R. (2002). Communities of practice with and without information technology. *Proceedings of ASIST, 39,* 338-349.

Kodama, M. (2001). New regional community creation, medical and educational applications through video-based networks. *Systems research and behavioral science, 18*, 225-240.

Laboratory for Intelligent Mechanical Systems, The. (2004). The Haptic community Web site. Retrieved March 1, 2004, from *http://haptic.mech.nwu.edu/*

Levine, A. & Sun, J.C. (2002). *Barriers to distance education*. Washington, DC: American Council on Education Center for Policy Analysis.

L-Soft. (2003). The history of LISTSERV. Retrieved July 30, 2003, from *http://www.lsoft.com/products/listserv-history.asp#lsoft*

Millen, D.R. & Dray, S.M. (2000). Information sharing in an online community of journalists. *Aslib Proceedings, 52*(5), 166-173.

Mirriam-Webster, Inc. (2004). *Community*. Retreived March 1, 2004, from *http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=community*

Nonnecke, B. & Preece, J. (2003). Silent participants: Getting to know lurkers better. In C. Leug & D. Fisher (Eds.), *From usenet to CoWebs: Interacting with social information spaces.* Amsterdam: Springer-Verlag.

Preece, J. (2000). *Online communities: Designing usability, supporting sociability.* Chichester, UK: John Wiley & Sons.

Preece, J., Maloney-Krichmar, D., & Abras C. (2003). History of online communities. In K. Christensen & D. Levison (Eds.), *Encyclopedia of community: From village to the virtual World* (pp. 1023-1027). Thousand Oaks, CA: Sage Publications.

Rheingold, H.. (1993). *The virtual community*. Retrieved March 1, 2004, from *http://www.rheingold.com/vc/book/intro.html*

Rheingold, H. (2000). *The virtual community: Homesteading on the electronic frontier*. Cambridge, MA: MIT Press. Retrieved March 1, 2004, from *http://www.rheingold.com/vc/book/*

Shenk, D. (1998). *Data smog: Surviving the information glut*. San Francisco: Harper Collins Publishers.

Shirky, C. (2003). A group is its own worst enemy. Retrieved July 30, 2003, from *http://shirky.com/writings/group_enemy.html*

Ultima Online (2004). ORIGIN - Ultima Online - Home Page. Retrieved March 1, 2004, from *http://www.uo.com/*

UNESCO (2001). *Teacher education through distance learning: Technology - Curriculum - Cost - Evaluation (Summary of case studies)*. Paris: UNESCO.

Utzinger, J., Tanner, M. & Singer, B.H. (2001). The Internet: A valuable tool for Roll Back Malaria. *Trends in Parasitology*, *17*(4), 159-161.

Weinreich, F. (1997). Establishing a point of view toward virtual communities. *Computer-Mediated Communication Magazine*, February. Retrieved March 2, 2004, from *http://www.december.com/cmc/mag/1997/feb/wein.html*

Wellman, B. (2001). Computer networks as social networks. *Science*, *293*(14), 2031-2034.

Wikipedia. (2003). Community. Retrieved March 1, 2004, from *http://en.wikipedia.org/wiki/Community*

Wilson, S.M. & Peterson, L.C. (2002). The anthropology of online communities. *Annual Review of Anthropology*, *31*, 449-67.

Wing of Madness Inc. (2004). The wing of madness: Depression information, news and support. Retrieved March 1, 2004, from *http://www.wingofmadness.com/about.htm/*

KEY TERMS

Avatar: A computer generated representation almost always graphical in nature, and sometimes a three-dimensional construct that represents the user/operator in the virtual world.

Computer Mediated Communication: Communication that is facilitated by the use of a computer, mainly through some form of technology such as a modem or digital connection to the Internet for the purposes of sending text, voice, or video to the recipient.

Haptics: The technology of touch which uses the tactile sense to send and receive data.

Information Technology: Encompasses all forms of technology used in processing and disseminating information.

Lurker: A member of an online community, discussion board, or Web site that interacts only passively.

MUD/MOO: Originally a multi-user dungeon, a virtual world where people would role play in a fantasy world. Later to expand to all manner of "virtual worlds".

Virtual/Online Community: A group of like-minded individuals who use computer mediated communication as a method of accessing or sending digital communication to each other. These individuals are usually related to each other through a common tie such as an information need, social bond, political bond, or any other bond. This term can be applied to any number of online entities such as chat rooms, message boards, Web sites, listservs, or other collaborative environments.

Principles for Managing Information Security

Rana Tassabehji
University of Bradford, UK

A BRIEF HISTORY

Information security has traditionally been the responsibility of information technology (IT) departments, where information security has commonly been perceived to have a technological solution. During the 1980s, computer usage was mainly concentrated in computer centres, where the implementation of computer security focused largely on securing the physical computer infrastructure of the organisation (Mutsaers et al., 1998), which proved highly effective.

The advent of cheaper and more powerful microprocessor, computing and networking technology dramatically changed the nature of computer usage in organisations. In the 1990s, the majority of employees had a workstation or personal computer through which they could directly access, process and manage any number of available corporate resources, such as software, hardware and information, to execute a wide range of tasks. In line with this, IT security developed additional technical measures that incorporated software residing on IT systems. These were able to deal with the increasingly new and varied attacks resulting from the wide use of distributed and inter-networked computers (Von Solms 1999; Vermeulen & Von Solms, 2002).Examples of these developments in IT security included user identification and authentication, memory clearance and access control to data.

The ubiquity of the Internet and e-mail has continued to increase the importance of information and related technology in organisations. In today's information economy, information is one of the most important assets for an organisation (Turner, 2000). Possession of strategic information can make the difference between an organisation's success and failure (Forcht, 1994). Not only information, but also the systems that support it, have become a critical part of an organisation's business and assets — a fact of which many attackers are well aware. According to a number of surveys (PricewaterhouseCoopers, 2002, 2004; Computing Technology Industry Association, 2003, 2004; Richardson, 2003), attacks on corporate information systems have been increasing year after year, with the costs of the security breaches in terms of disruption, damage and loss also increasing. These surveys are probably an underestimation of the total number of security breaches that have occurred, since breaches can go undetected or are unreported and not publicised for fear of repercussions (negative publicity or lawsuits). One of the cardinal rules for effective management of security is to say nothing about security. So with the prominence of information also comes the heightened importance of securing information and managing the security process.

THE INTERNET AGE

As well as being a core business asset, information systems are increasingly recognised as socio-technical infrastructures that rely heavily on people. In recent years, it has become more widely acknowledged that human factors play a part in many security failures (Weirich & Sasse, 2002). As such, managing information security has begun to move out of the technology department as the sole source of responsibility and solutions to a more holistic organisational approach that incorporates business processes, controls and policies; corporate governance; systems and technology infrastructures; human resource management and training; and organisational culture (Higgins, 1999; Gelbsein, 2001; Eloff & Eloff, 2003; Tassabehji, 2003). Empirical evidence supports this view: a survey by McKinsey (McKinsey Quarterly, 2002) found that although only 6% of Fortune 500 companies had appointed a senior business executive to oversee information security, this figure was expected to rise over the next few years, as strategic, operational and organisational safeguards are added to the technological measures being employed to protect corpo-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

rate information. One of the main catalysts that has mobilised organisations to think more seriously about information security is the introduction of new and modified legislation (such as the Data Protection and other Electronic Communications Acts). There is now international legal recognition of the importance of information and the need to secure it. Organisations must ensure that data is protected as the legal responsibility for protection falls fully on the organisations that collect, store, share and use the data.

THEORY

There is very little academic theory that deals solely with managing information security. The majority of publications on the topic are largely practitioner based, relying on standards, benchmarks, best practice, technical specifications, security frameworks and models. Hong et al. (2003) identify five theories that define approaches to the management of information security:

- **Security policy theory:** aims at establishing, implementing and maintaining an organisation's information security requirements through a security policy.
- **Risk management theory:** evaluates and analyses threats and vulnerabilities of information assets in an organisation. It also includes the establishment and implementation of control measures and procedures to minimise risk.
- **Control and audit theory:** suggests that organisations should establish control systems (in the form of security strategies and standards) with regular auditing to measure control performance.
- **Management system theory:** establishes and maintains a documented information security management system. This includes an information security policy that incorporates factors internal and external to the organisation; the scope of the policy; risk management and implementation of the process.
- **Contingency theory:** information security is a part of contingency management that prevents, detects and reacts to threats and vulnerabilities internal and external to an organisation. This incorporates all the other approaches identified above in order to manage the threats.

Although these are presented as separate theories or approaches, they are not mutually exclusive. Each of the above approaches incorporates one or more of the other elements but from different perspectives that emphasise specific issues. Hong et al. (2003) highlight some limitations of each of the theories — for instance, all except contingency theory take a top-down approach, which may not be consistent with reality. They posit an "integrated system theory" based on contingency management and which also integrates information security policy, risk management, internal control and information auditing theories to form an Information Security Architecture consistent with organisational objectives. However, this integrated theory does not detail the measures under which each of the approaches functions or interacts with the other. It is an outline framework that attempts to take a holistic approach to information security management in the same vein as other information security academics such as Eloff (1998), Van Solms (1999) and Higgins (1999) but focusing on the development of a more "theoretical" framework.

THE MAIN COMPONENTS

Whatever the emphasis that an organisation places on managing information security, the main components that must be included are organisation and IT infrastructure; risk assessment and management; security policy, standards and procedures; security awareness and training programmes; monitoring, auditing, reviewing; and updating policies and processes.

Each of these five major components is linked together in a cyclical feedback process, where they are all interdependent and as a holistic process contribute to the overall security of the organisation's information. Each of the five components will be discussed in detail in the following sections.

Organisation and IT Infrastructure

It is widely recognised in management theory that commitment and support from top management is

Figure 1. Main components for managing information security

critical to the successful implementation of new technology, systems or processes in organisations (Beaver, 2003; Edmondson, 2003; Hansson et al., 2003; O'Neil & Baker, 2003; Sarker & Lee, 2003; Sharma & Yetton, 2003; Taylor & Wright, 2003). This is the same for implementing information security. In addition, there is also a need to align security and business objectives and activities (Von Solms, 1998; Kwok & Longley, 1999) to ensure that the security service being implemented is actually addressing business issues and objectives.

Risk Assessment and Management

Managing systems risk is based on the concept of deterrence theory, which highlights (1) deterrence, (2) prevention, (3) detection, and (4) recovery. Straub and his research partners (Goodhue & Straub, 1991; Straub & Welke, 1998) have applied deterrence theory to the IS environment, which posits that information security actions can deter potential computer abusers from committing acts that implicitly or explicitly violate organisational policy. This premise is supported by empirical evidence in their research, which shows that security actions can lower systems risk but that managers often lack this security knowledge. Thus, their subsequent actions to cope with systems risk are less effective than they might otherwise be. Straub et al.'s (1998) research highlights the need for an overall security risk planning process, which is lacking in the public domain. This would constitute five stages:

a. Recognition of security problems.
b. Risk analysis, including threat identification and prioritisation of risks.
c. Alternatives generation to identify solutions to meet the organisational needs in a specific risk situation.
d. Decisions matching threats with appropriate solutions.
e. Implementation of the realised plans.

Bruce Schneier (Heiser, 2003) similarly identifies a five-step process for analysing specific risk decisions:

a. Identifying the assets needing protection.
b. Analysing the risk.
c. Evaluating the effectiveness of the proposed solution.
d. Identifying any other risks caused by the proposed solutions.
e. Understanding the inherent trade-offs.

The techniques for performing risk assessment are varied, ranging from the more traditional mathematically based checklists and probability risk analysis to less mathematically exact threat tree analysis based on matches to semantic terms such as "high," "low" and "moderate" risk as recommended by the United States (U.S.) Department of Defense (Straub & Welke, 1998). The ultimate aim of any risk assessment process is to prioritise and then determine the acceptable level of risk for each individual organisation. However, traditional risk analysis methods cannot adequately reflect the loss from the disruption of operations resulting from asset failure (Suh & Han, 2003). Suh and Han (2003) developed an IS risk analysis methodology based on a business model that considers the replacement of assets and the disruption of operations, as well as the value of assets based on the relationship between assets and business functions. The value of assets reflects the value of the business functions that the assets support. There are three main steps to this process — the organisational investigation; asset identification and evaluation; threat and vulnerability assessment — all of which

are similar to the previous methods except that an additional organisational dimension is included. So the foundations of risk assessment have been laid down, and subsequent refinements ensure that risks are being assessed accurately and according to the value placed by each respective organisation.

Security Policy, Standards and Procedures

A crucial part of managing information security is having a framework and set of standards to which all the necessary areas of information security in the organisation adhere. A number of different international standards and best practice guidelines exist. Many of these underline the same areas of importance to be addressed, and the majority use the British Standard BS7799 (now ISO 17799) as their foundation.

Despite the fact that there are many national and international standards governing the management of information security (summarised in the key terms section), there are many common areas between them. For instance, need and scope of information security, management commitment, asset and risk assessment. None of the standards are prescriptive about the content of the policy documentation, neither do they include a comprehensive discussion of information security policy (Hone & Eloff, 2002). The respective sets of standards attempt to describe the various processes and controls needed for successfully implementing an information security policy, rather than advising what should be included in the document; "international standards will not write the document for you" (Hone & Eloff, 2002, p. 409). However, since the organisational security policy is one of the most important documents, it must be written not only to abide by the standards stated, but also to ensure that the context and background of the organisation is included.

Rees et al. (2003) developed a policy framework for information security. They maintain that the problems with security policies are that they tend to be technology neutral; set direction and procedures; and define penalties and countermeasures if the policy is transgressed. They posit that security policies need to keep up with both the rapid changes in technology and also the changes in organisational objectives. They developed a tool called FIRES[1] to aid in the formulation and management of security policies, which draws on the new product development cycle and the systems development life cycle. They see the FIRES lifecycle as an iterative process, with feedback loops at every step, consisting of four major stages: *Assess*, reviewing existing policies, standards, guidelines and procedures; *Plan*, preparing for the implementation of the proposed change(s), which include policy development and requirements definition; *Deliver*, implementing the controls defined; and finally, *Operate*, a daily phase where controls are complied with, monitored, managed and reviewed. How easy it is to implement such a dynamic tool has yet to be proven; however, that developers have aimed to mirror the rapidly changing nature of products and services in organisations as well as the technological infrastructure that supports them is a valid premise and one which chimes with management and technological security requirements. More research needs to be conducted in this area to further validate this model.

However, despite the introduction of numerous models for developing and implementing policies, one main issue remains: unless the policy is implemented by the people in the organisations, information and systems will still remain at risk. Policies must be user-centred (Besnard & Arief, 2004), relying on the rules of human-computer interaction, which ensure maximum engagement by all users. More empirical and practitioner-based research into making user-centric security policies that can be remembered and applied by legitimate users is needed, as this is one area that is still research young.

Security Awareness and Training

Training and raising awareness amongst organisational workers is advocated in information security literature (Straub & Welke, 1998; Eloff & Eloff, 2003) and is a large element of the majority of the published guidelines and standards. Besnard and Arief (2004) include an additional dimension to understanding information security training by using research in psychology to highlight that human beings make biased decisions and, as such, obey least-effort rules where they subjectively weigh up maximum benefits for the cost of a given action at the

expense of rules. This then leads to *"insidious security lapses, where the level of protection is traded off against usability"* (Besnard & Arief, 2004). One of the main recommendations is to educate staff in a way that will at least allow them to be aware of the consequences of insecure practices and that they will distinctly remember. Research still needs to be done in this area to evaluate how and whether different or new techniques are needed specifically for information security-based training.

Monitoring, Reviewing and Auditing

Monitoring, reviewing and auditing are processes that have become embedded in quality assurance and best practice throughout much organisational and business management and planning. The security standards, policies and procedures underline the need for monitoring, reviewing and auditing the information security infrastructure that has been implemented. Information security is a fast moving area, and as such, the infrastructure developed to protect it must also be constantly upgraded and updated.

CONCLUSION

Information has become a critical part of socio-economic life in the 21st century, where it is one of the most valuable assets any organisation holds. Securing information is an equally critical part of an organisation's operational, managerial and strategic function. Information security management is no longer solely about technology, but includes a wide spectrum of issues to be addressed during planning, management and monitoring of information security within an organisation. Information security management encompasses managerial, technical, strategic (corporate governance, policies and pure management) and human (culture, training, ethics, awareness) issues. Managing information security should be based on business needs, and its management should exhibit an integrated, holistic, organisational approach where the implementation of, and compliance with, controls and procedures can be developed according to published standards or codes of practice (e.g., ISO17799) (Vermeulen & Von Solms, 2002).

Organisations are beginning to take information security seriously, so business is increasingly beginning to realise that there are also organisational, operational and strategic responsibilities and solutions for information security. More research into organisations' management of information security is needed to consolidate best practice. However, empirical research into security is a field in which it is notoriously difficult to attract organisations to engage (Kotulic & Clark, 2004). So any future research must incorporate a methodology that takes on an added dimension of sensitivity and time in which trust can be built.

REFERENCES

Beaver, G. (2003). Successful strategic change: some managerial guidelines. *Strategic Change, 12*(7), 345.

Besnard, D., & Arief, B. (2004). Computer security impaired by legitimate users. *Computers & security, 23*(3), 253-264.

CompTIA. (2003). *Committing to security: A CompTIA analysis of IT security and the workforce.* Retrieved March 24, 2004, from *www.comptia.org/pressroom/get_news_item.asp?id=364*

CompTIA. (2004). Computer viruses, worms pose biggest security headache for IT departments. WebPoll, retrieved March 24, 2004, from *www.comptia.org/pressroom/get_news_ item.asp ?id=364*

Edmondson, A.C. (2003). Framing for learning: Lessons in successful technology implementation. *California Management Review, 45*(2), 34.

Eloff, J., & Eloff, M. (2003). Information security management – A new paradigm. *Proceedings of the 2003 SAICSIT.* ACM International Conference Proceeding Series.

Forcht, K.A. (1994). *Computer security management*. Boyd & Fraser.

Gelbsein, E. (2001). *Managing information security.* OECD Workshop, International Computing Centre, Geneva. Retrieved February 20, 2004, from

http://accsubs.unsystem.org/ccaqfb-intranet/Productivity/IT/Managing%20information %20security %20OECD.pdf

Goodhue, D.L., & Straub, D.W. (1991). Security concerns of system users: A study of perceptions of the adequacy of security measures. *Information & Management, 20*(1), 13-27.

Hansson, J. Backlund, F., & Lycke, L. (2003). Managing commitment: increasing the odds for successful implementation of TQM, TPM or RCM. *The International Journal of Quality & Reliability Management, 20*(8/9), 993.

Heiser, J.G. (2003). Beyond cryptography: Bruce Schneier's beyond fear: thinking sensibly about security in an uncertain way. *Computers & Security, 22*(8), 673-674.

Higgins, H.N. (1999). Corporate system security: Towards an integrated management approach. *Information Management & Computer Security, 7*(5), 217.

Hone, K., & Eloff, J.H.P. (2002). Information security policy – What do international information security standards say? *Computers & Security, 21*(5), 402-409.

Hong, K., Chi, Y., Chao, L.R., & Tang, J. (2003). An integrated system theory of information security management. *Information Management & Computer Security, 11*(5), 243-248.

Kwok, L., & Longley, D. (1999). Information security management and modelling. *Information Management & Computer Security, 7*(1), 30-39.

McKinsey Quarterly. (2002). Managing information security. Retrieved March 24, 2004, from *http://news.com.com/2009-1017-933185.html*

Mutsaers, E.-J., Zee, H.V.D., & Giertz, H. (1998). The evolution of information technology. *Information Management & Computer Security, 6*(3), 115.

O'Neil, D.V., & Baker, P.M.A. (2003). The role of institutional motivations in technological adoption. *Information Society, 19*(4), 305.

PricewaterhouseCoopers. (2002). Department of Trade and Industry: Information security breaches survey. Retrieved March 24, 2004 from *http://www.dti.gov.uk/industries/information_security/downloads.html*

PricewaterhouseCoopers. (2004). Department of Trade and Industry: Information security breaches survey. Retrieved March 24, 2004 from *http://www.dti.gov.uk/industries/information_security/downloads.html*

Rees, J., Bandyopadhyay, S., & Spafford, E. (2003). FPIRES: A policy framework for information security. *Communications of the ACM, 46*(7), 101-106.

Richardson, R. (2003). *Eighth annual CSI & FBI computer crime and security survey.* San Francisco: Computer Security Institute.

Sarker, S., & Lee, A.S. (2003). Using a case study to test the role of three key social enablers in ERP implementation. *Information & Management, 40*(8), 813.

Sharma, R., & Yetton, P. (2003). The contingent effects of management support and task interdependence on successful information system implementation. *MIS Quarterly, 27*(4), 533.

Straub, D.W., & Welke, R.J. (1998). Coping with systems risk: Security planning models for management decision making. *MIS Quarterly, 22*(4), 441.

Suh, B., & Han, I. (2003). The IS risk analysis based on a business model. *Information and Management, 41*(2), 149-158.

Tassabehji, R. (2003). *Applying e-commerce in business.* SAGE.

Taylor, W.A., & Wright, G.H. (2003). A longitudinal study of TQM implementation: Factors influencing success and failure. *Omega, 31*(2), 97.

Turner, C. (2000). *The information e-conomy*. Kogan Page.

Vermeulen, C., & Von Solms, R. (2002). The information security management toolbox – Taking the pain out of security management. *Information Management & Computer Security, 10*(3), 119-125.

Von Solms, R. (1998). Information security management (3): The code of practice for information security management (BS7799). *Information Management & Computer Security, 6*(5), 224-225.

Von Solms, R. (1999). Information security management: why standards are important information. *Information Management & Computer Security, 7*(1), 50-57.

Weirich, D., & Sasse, M.A. (2002). *Pretty Good Persuasion: A first step towards effective password security in the real world*. ACM/SIGSAC New Security Paradigms Workshop, New Mexico.

KEY TERMS

BS7799-2:2002: Part 2 is an Information Security Management System (ISMS) that adopts a systematic approach to managing sensitive company information that encompasses people, processes and IT systems. This is under revision and is expected to be complete in late 2004/early 2005.

Control Objectives for Information and Related Technology (COBIT): Designed to be IT governance aid for understanding and managing the risks and benefits associated with information and related technology. It is intended that CobiT provide clear policy and good practice for IT governance throughout the organisation.

Generally Accepted System Security Principles (GASSP): Developed by the U.S. National Research Council, it considers the terms policy, rules, procedures and practices that relate to organisational implementation of physical, technical and administrative information security. It incorporates the consensus as to accepted information security principles of the GASSP Committee, followed by international IT community review at a particular point in time. (http://web.mit.edu/security/www/GASSP/gassp021.html)

Guidelines for the Management of IT Security (GMITS): *(ISO 13335*) A five-part series of technical reports providing guidance on the management of IT security to assist organisations in developing and enhancing their internal security architecture, and as a means to establish commonality between organisations. (www.it-security.sk/iso_13335_an.htm)

International Standards Organisation (ISO): A non-governmental organisation constituting a network of the national standards institutes of 148 countries who work together to develop standards for traditional activities, such as agriculture and construction, to the newest information technology developments, such as the digital coding of audio-visual signals for multimedia applications. "*Without the international agreement contained in ISO standards on quantities and units, shopping and trade would be haphazard, science would be – unscientific – and technological development would be handicapped."* (www.iso.org)

ISO/IEC 17799: The British Standards Institute published a code of practice for managing information security following consultation with leading companies. BS7799 Part 1 incorporates a broad range of security practices and procedures that can be adopted by any organisation of any size and in any industry sector. This is now international standard.

Operationally Critical Threat Asset and Vulnerability Evaluation (OCTAVE): OCTAVE provided details of accepted best practices for evaluating security programmes.

ENDNOTE

[1] Policy Framework for Interpreting Risk in E-business Security

Privilege Management Infrastructure

Darren P. Mundy
University of Hull, UK

Oleksandr Otenko
University of Kent, UK

PRIVILEGE MANAGEMENT INFRASTRUCTURE: AN OVERVIEW

Public Key Infrastructures (PKI) are now in place in a number of organizations, and there is a wide amount of material available that can be used to obtain familiarisation with the concept (Adams & Lloyd, 2002; Housley & Polk, 2001; Nash, Brink, & Duane, 2001). Although related to PKI, Privilege Management Infrastructure (PMI) is a recent development in the network security field. PMI has been designed to supply the authorization function lacking in the PKI model. This article will provide an overview of PMI, state the relationship between PKI and PMI, and will finally provide a number of examples of present PMI architectures such as PERMIS (Chadwick & Otenko, 2002), OASIS Active Security (Yao, Moody, & Bacon, 2001) and AKENTI (Thompson, Essari, & Mudumbai, 2003).

WHAT IS PMI?

PMI can generally be thought of as the infrastructure supporting a strong authorization subsystem via the management and use of privileges (Adams & Lloyd, 2002). PMI is essentially a term used to encompass the management of authorization processes such as access control, rights management, levels of authority, delegation of authority, and so on. A PMI helps an organization to provide secure access to any target resource they specify based on policy. A policy should detail such information as which users are allowed access to which resources, what actions they are allowed to perform, when they are allowed access, for example, time constraints, what privileges they need to be able to access the resource and carry out an operation.

As organizations embrace electronic business they are increasingly striving to provide electronic access to greater amounts of organizational resources to improve their services and decrease transaction costs. However, by opening up electronic access to their resources for their partners, clients, employees, and so on … they are heightening the security risks that they face (Newman, 2003). Organizations need to be sure that access to their resources is controlled by a variety of security mechanisms, for example:

1. To ensure the party requesting access is who they say they are (authentication).
2. That the party has sufficient rights to access the resource (authorization).
3. That confidential material is only read by those authenticated and authorized parties (privacy).
4. That the transaction is monitored (audit and control).

PMI addresses only authorization. To address other points, corresponding subsystems should be deployed. In further sections it will be explained how the PMIs and PKIs are related, and examples of the use of PMIs will be provided. The discrepancies between various implementations will be highlighted, and the difficulties with using PMIs will be discussed.

The Relationship Between PKI And PMI

The authorization subsystem supported by a PMI can be relied upon to control access based on the privileges possessed by a user. However, it doesn't provide any assurance as to the user's identity. To ensure identity we require an identity management system such as a PKI. Familiarity with the PKI concept can be obtained using the wealth of available literature

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

(Adams & Lloyd, 2002; Housley & Polk, 2001; Nash et al., 2001). In this context however it can simply be stated that a PKI provides authentication services whilst the PMI provides authorization services. There is a large similarity between PKI and PMI at multiple levels (Chadwick & Otenko, 2002), for example.

- In a PKI users are given digital certificates proving their identity; in a PMI users are given digital certificates proving their privilege(s).
- In a PKI a Certificate Authority issues the digital certificates; in a PMI an Attribute Authority creates the digital certificates.

PMI ARCHITECTURES FOR TRUST ESTABLISHMENT

In this section, Privilege Management will be looked at from the point of view of the access control system.

Prior to the introduction of Privilege Management Infrastructures (PMI), access control systems trust only their "local" information about the outer world. This is very effective for small groups of people (e.g., multi-user Operating Systems). However, when the number of users willing to co-operate increases, it becomes more difficult to reflect all of the circumstances of the world locally. Dynamicity of relationships between the resource owner and the users accessing the resource also increases the difficulty of managing the privileges each of the users has, limiting scalability of such systems.

To facilitate scalable solutions, trust in the people must be established in a distributed manner, and a means of distributing trust is required. This can be achieved in a number of ways. This section describes how this is done in three different PMI models. It starts with the approach adopted by X.509 and is followed by description of the Akenti and Active Security PMI architectures.

X.509

In X.509 there is a single root of trust, the Source of Authority (SOA). It stands for the owner of a resource, or an agent acting on his behalf. The SOA specifies the rules for establishing trust relationships, and access control rules. All such rules are written in a form of a policy, which governs the access control system. The SOA is also the ultimate authority in assigning privileges to end-entities, which will use the resource.

The SOA distributes the privilege to assign privileges to other entities, which are called Attribute Authorities (AAs), and the process of assigning this privilege is called delegation. These authorities, in their turn, may be allowed to assign privileges to end-entities, or delegate them further to other Attribute Authorities. Thus the PMI forms a tree of authorities, with a singular root, which is the Source of Authority. The leaf nodes are end-users, who can only assert their privileges, and cannot delegate them to other entities.

The fact of assignment of privilege to an entity (either to AA or to an end-user) is noted as an X.509 Attribute Certificate (AC), which is a digitally-signed document, describing who has assigned what privilege to what entity. The privilege in such ACs is specified in a form of a privilege attribute that has to be interpreted by the access control system.

The access control system can discover trust relationships between the SOA (the resource owner) and the end-entities by obtaining their ACs and validating their contents using the policy written by the SOA. To achieve this, the access control system must obtain the ACs of the end-entity attempting access, the ACs of the Authority assigning the privilege to it (remember, that X.509 ACs specify who the grantor was), and ACs of all AAs that granted the privilege to do this to the authority, and to each of those AAs. Then the system needs to validate each of the assignments that occurred against the policy: if so, then the end-entity has been assigned a privilege in a trustworthy way, and an access control decision can be made; if the assignment of some privilege is not allowed by the policy, the privilege assignment is not trustworthy and should be discounted when making an access control decision.

In X.509, privilege assignment is valid if the granted privilege is a subset of all the privileges the grantor has, the only exception being the source of authority, which can assign any privilege to any entity[1]. To be able to make judgments if a granted set of privileges is a subset of the privileges the grantor had, the privilege attribute values must have order. Some access control models (MAC, RBAC) naturally have ordering of privilege attribute values; other models may need enhancement (Otenko, 2004).

To summarise, X.509 PMIs form a tree with the root in the source of authority, which is essentially the owner or the governor of the resource. Trust relationships are established by issuing X.509 ACs with privilege attributes to entities. The SOA writes a policy to which all the trust relationships must conform. The rules for validating privilege assignments ensure that trust does not increase when it is distributed down the PMI tree away from the SOA. PERMIS is one of the implementations of X.509 PMI with Role Based Access Controls (Chadwick & Otenko, 2002).

Akenti

Akenti is an access control system that has been developed by Lawrence Berkeley National Laboratory, USA (Thompson, Essari, & Mudumbai, 2003). It implements a model of PMI that is somewhat different from the (traditional) X.509 model.

In this model, like in X.509, there are Attribute Authorities that assign privileges to other AAs and end-entities[2], thus the PMI also forms a tree. However, there is no singular root, there is no SOA in their model. Instead, Stakeholders are introduced, who jointly govern the resource (they may be responsible for different aspects of access control). Thus the Akenti PMI becomes a multi-root tree.

The resource is governed by a policy issued by some stakeholder. Even though only one Stakeholder can issue such policy, any of them can do this, and even after the policy has been written, any stakeholder can issue supplementary statements in a form of use condition certificates, which can restrict or enhance the (original) policy. It is because stakeholders are independent sources of trust that we can talk about them as multiple roots of the PMI.

The access control system discovers what privilege an end-entity has in a way similar to the X.509 standard, only that there are several equal roots of the PMI.

To summarise, the Akenti PMI has multiple roots, and this is the main difference between this PMI and the (traditional) X.509 PMI. This solution makes the administration of the resource more flexible (as compared to the X.509 model), but may cause unexpected access control decisions, if the stakeholders (the roots of the PMI) create contradictory statements, when they operate unaware of each other's requirements. Further differences between the Akenti PMI and the PERMIS X.509 PMI can be found detailed in Chadwick and Otenko (2003).

OASIS Active Security

The OASIS (Open Architecture for Secure Interworking Services) research group at Cambridge University have developed a different PMI architecture. In their model, many novel concepts are used, which are very efficient in dynamic environments. In particular, active security has an updated definition of delegation, which is called appointments. The assignment of privileges is done to perform particular duty, so the grantor of the privilege does not necessarily have the privilege he gives to others, (e.g., a project manager) does not need to have permission to perform a specialised operation, but he can appoint a competent person to do that.

The researchers argue that such a model is more adequate than the X.509 model for many reasons. This concept, as Yao et al. (2001) point out, can also be implemented by using different hierarchical relationships specified between the roles. One of the hierarchies would be the specification of the privilege inheritance properties, the other would be defining what roles can appoint what other roles to end-users.

To summarise, the active security model proposes a different approach to assigning privileges. The use of the appointments instead of delegation requires more active interoperation between the privilege issuers and the privilege verifiers, than this is required for X.509 PMIs.

ACTUAL EXAMPLES OF PMI USAGE

In this section a number of real life examples where PMI has been used to provide authorization management are detailed, in the areas of electronic governance and health care. Further practical examples of PMI usage in practice can be found in Jansen (2001), Chadwick (2003), Jin, Kim, and Ryou (2002), and Fagotto, Ferrer-Roca, Espinosa, Suarez, and de Leon (2001).

Electronic-Governance

An X.509 role based PMI has been used across Europe to provide electronic government services in

the cities of Barcelona (Spain), Salford (UK) and Bologna (Italy) (Chadwick & Otenko, 2002). Each of the cities already had experience in running PKI services, and wished to extend their security infrastructures with a PMI implementation to create a strong authentication and authorization chain. The applications in which PMI has been applied differ for each municipality.

Barcelona required a PMI to allow electronic access to the city parking fines database for car-hire organizations. Allowing such access means the car-hire companies can automatically check when a car is returned that no parking tickets have been issued to the customer. If a parking ticket has been issued, then the car-hire company can transfer the fine to the individual. A council SOA assigns a role value of either generalised or authorised. Any citizen or business can be allocated the generalised role. Anyone with the generalised role has permission to read their own pending car parking fines. Businesses that have signed an agreement with the Barcelona city council are given the authorised role. Authorised roles can read their own pending fines and also may modify the details of them (e.g., update the driver's name and address).

The Salford City Council has used a PMI to control authorization in an electronic tendering application. The city council generates Request for Proposal (RFP) documents allowing anyone to download them. However, for a number of proposals only companies registered with the council or only companies possessing a certain requirement (e.g., ISO 9000 certification) are able to submit tenders. The tenders remain anonymous until the winner has been chosen. In Salford's case, a council SOA may assign a role value of either tenderer or tender-officer, whilst an external SOA may assign ISOCertified roles to companies. The city tender policy does not allow access to the electronic tender store before the closing date of the RFP, and tenderers are not allowed to submit tenders after the closing date of the RFP.

Finally, Bologna has constructed a similar application that utilizes PMI to provide authorization to a city-planning server.

Health Care

A PMI infrastructure has been used to manage the authorization infrastructure in the electronic transfer of medicinal prescriptions (ETP) for the UK National Health Service (Chadwick & Mundy, 2003). The assignment of roles is distributed to the appropriate authorities in the health care and government sectors. This includes the assignment of both professional roles such as doctor and dentist, as well as patient roles that entitle patients to exemption from payment for a variety of reasons for example, claiming social security benefits and children under 16, and so on. All roles are stored as X.509 ACs in LDAP directories, which are managed by the assigning authorities. The PERMIS policy based decision engine subsequently retrieves these ACs containing roles, in order to make granted or denied access control decisions, which are required by the prescribing and dispensing applications. The SOA for setting the ETP policy is assumed to be the UK's secretary of state for health. The ETP policy says what roles are recognized, who is authorized to assign the roles, what privileges are granted to each role and what conditions are attached to these privileges. The ETP policy is then formatted in XML, embedded in an X.509 attribute certificate, digitally signed by the secretary of state for health, and finally stored in an LDAP directory. From here it can be accessed by all the ETP applications in the UK National Health Service that contain embedded policy based PERMIS decision engines.

CRITICAL ISSUES IN PMI USAGE

There are a number of key issues that organizations need to consider before the implementation of any PMI system. These include:

- Security policies and practices should be defined in clear, unambiguous statements that cover any possible circumstances
- Present Security Infrastructure: PMI builds on the authentication infrastructure provided by a PKI to create a strong authorization and authentication chain. In many scenarios authentication and authorization are required therefore it is reasonable to assume that any organization wishing to implement a PMI system should also have in place (or be looking to implement) a secure authentication system.

- Administration: Dependant on the application sub-context PMI may require a significant administration load on the organization.
- Privilege Storage & Retrieval: Organizations need to consider how and where the privileges are to be stored. For example, a denial of service attack on a single storage resource may result in PMI failure.
- Granularity of Privileges: Privileges can be granted at a number of different levels, for example, job role through to single task authorization. Organizations need to carefully consider the granularity of their PMI as this can have a severe impact on the administration of the system.

CONCLUSION

The move towards widespread remote electronic access to organizational resources brings with it severe security concerns especially in the present environment with security threats ever increasing (CSI, 2003). PMI is a new concept in the field of Information Security Management specifically designed to solve the authorization element of secure electronic access to data. The level of protection is increased even further with the combination of an authentication management infrastructure (such as a PKI) with PMI (Dawson, Lopez, Montenegro, & Okamoto, 2002). The use of PMI will continue to grow as PMI mechanisms are adapted to new environments, such as in the Grid applications (Chadwick, 2003) and mobile agents (Jansen, 2001).

REFERENCES

Adams, C. & Lloyd, S. (2002). *Understanding public-key infrastructure: Concepts, standards, and deployment considerations* (2nd ed.). New York: Macmillan.

Chadwick, D.W. (2003). *An authorisation interface for the GRID.* Presented at the E-Science All Hands Meeting, Nottingham, UK.

Chadwick, D.W. & Mundy, D. (2003). Policy based electronic transmission of prescriptions. *Procedings of the Fourth IEEE International Workshop on Policies for Distributed Systems and Networks*, Lake Como, Italy, (pp. 197-206).

Chadwick, D.W. & Otenko, A. (2002). The PERMIS X.509 Role Based Privilege Management Infrastructure. *Future generation computer systems, Elsevier Science BV.*, 936, 1-13.

Chadwick D.W. & Otenko, O. (2003). A comparison of the Akenti and PERMIS authorization infrastructures, in ensuring security in IT infrastructures. In M.T. El-Hadidi (Ed.), *Proceedings of the ITI First International Conference on Information and Communications Technology*, Cairo University, (pp. 5-26).

CSI/FBI Computer Crime and Security Survey (2003). Richardson, R. (Editorial Director), Computer Security Research Institute.

Dawson, E., Lopez, J., Montenegro, J.A., & Okamoto, E. (2002). A new design of privilege management infrastructure (PMIs) for organizations using outsourced PKI. *The Fifth International Conference on Information Security,* LNCS 2433, Springer-Verlag, Sao Paulo, Brazil), 136-149.

Fagotto, F., Ferrer-Roca, O., Espinosa, Y., Suarez, M., & de Leon, J. (2001). Attribute certificates in e-health privilege management infrastructure. *Proceedings of the Sixth World Congress on the Internet in Medicine*, published as a special issue of Technology and Healthcare. Amsterdam: IOS Press.

Housley, R. & Polk, T. (2001). *Planning for PKI: Best practices guide for deploying public key infrastructure.* New York: Wiley.

Jansen, W. (2001). Determining privileges of mobile agents. *Proceedings of the Computer Security Applications Conference.*

Jin, S., Kim, H., & Ryou, J-C. (2002). Global public key infrastructure for secure e-commerce. *Proceedings of the Second International Workshop for Asia Public Key Infrastructure*, Taipei, Taiwan.

Nash, A., Brink, D., & Duane, B. (2001). *Implementing and managing e-security.* Osbourne McGraw-Hill.

Newman (2003). *Enterprise security*. Prentice Hall.

Otenko, O. (2004). *Policy-based privilege management using X.509 (The PERMIS Project)*. PhD Thesis, University of Salford, UK

Thompson, M., Essari, S., & Mudumbai, S. (2003). Certificate-based authorization policy in a PKI environment. *ACM Transactions on Information and System Security*, *6*(4), 566-588.

Yao, W., Moody, K., & Bacon, J. (2001). A model of OASIS role-based access control and its support for active security, *SACMAT'01*, Virginia, USA

KEY TERMS

Access Control: Restriction of access to some resource through the application of a mechanism which grants, denies or revokes permissions.

Access Rights Management: The process of assigning digital rights to users which can then be used in conjunction with an access control system to obtain access to some resource. The management infrastructure covers for example the allocation, renewal, and revocation of users rights.

Authentication: Is the process by which a system can provably verify the identity of a resource such as an individual, a system, an application, and so on.

Authorization: The process of determining if a requesting party has sufficient rights to access a resource.

Certificate Authority: An authority that manages the allocation of digital identity certificates to users. The CA exists as part of a PKI. The CA in conjunction with a Registration Authority (RA) initially checks to ensure the identity of a user. Once identity has been confirmed, the CA issues digital identity certificates that electronically assure the identity of a user based on the CA's digital signature.

Digital Certificate: A digital certificate is an electronic "passport", which can be used to establish identity in an electronic environment.

Digital Signature: An electronic signature can be deemed the digital equivalent of a handwritten signature. Electronic signatures can be used to authenticate the identity of the signer of the document and to also confirm the data integrity of the document.

Privilege: In a PMI, a privilege can be defined as an electronic right given to users enabling them to access various resources.

Privilege Delegation: The process by which a privilege given to one party can be transferred to another party either for an indeterminate or definite period of time.

ENDNOTES

1 The other way of looking at this is that the SOA has the whole universe of privileges within its PMI, so any privilege is a subset of the SOA's privileges anyway.

2 The format of the ACs is proprietary, not X.509.

Production, Delivery and Playback of 3D Graphics

Thomas Di Giacomo
University of Geneva, Switzerland

Chris Joslin
University of Geneva, Switzerland

Nadia Magnenat-Thalmann
University of Geneva, Switzerland

INTRODUCTION

Three-dimensional (3D) representation is one of the cornerstones of Computer Graphics (CG) and multimedia content. Advances in this domain, coupled with the highly fuelled progression of 3D graphics cards, have pushed the complexity of these representations into a whole new era, whereby a single real-time model can consist of more than a million polygons. Huge architectural buildings, everyday objects, even humans themselves, can be represented using 3D graphics in such detail that it is difficult to distinguish between real and virtual objects. Concurrently, and much towards the other end of the scale, many devices, such as Personal Digital Assistants (PDAs), mobile phones, laptops and so forth, are now "3D capable" to enhance a user's experience and to provide much more depth to the information presented. In many cases, these devices access the same content from the same service provider; for example, providing virtual maps/guides, multi-user games and so forth. It is this broadness of content and the heterogeneity of devices in terms of performance, capability, network connection and more that is the main concern in a continuously expanding market. It is also the concern of users to obtain the best quality for their device; that is, the general expectation of any device of higher performance is that overall the quality of the experience will be better.

Overall, three main stages have to be ensured to meet such requirements within an entire integrated chain. First, 3D media contents, from 3D models to animation data, have to be designed and produced by content providers, designers or artists, for instance. Second is to deliver all these pieces of content at a user's request, and although this kind of data has mainly been stored locally on the system, it is now more likely to be delivered via networks (either via download or streaming), similar to other media types (such as video and audio). Finally, upon reception, this media content has to be consumed by the users; that is, played back on their device. After a brief discussion on the background and concepts required for such goals, those three main steps are presented in detail in the following respective sections.

BACKGROUND

Though 3D technology is often considered, by misconception, as a local storage of data accessed by stand-alone applications – for example, video games, as Joslin, Di Giacomo and Magnenat-Thalmann. (2004) discuss – collaborative virtual environments have opened the way for distributed and networked 3D applications. Such architectures are becoming more and more common, and today's Web graphics are very much evolved. Lau, Li, Kunii, Guo, Zhang, Magnenat-Thalmann, Kshirsagar, Thalmann and Gutierrez (2003) present the emergence of standards that provide generic tools and formats for an even wider availability of such systems. Furthermore, standards enable interoperability and genericity, and extend the usability of graphics-based distributed applications. The MPEG group (see Walsh & Bourges-Sévenier, 2001 for a detailed description on MPEG-4) is one of the most important actors of these standardization efforts. Though it handles many dif-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

ferent types of media, and while originally the most commonly used were video and audio, 3D graphics is now receiving a great deal of attention, especially in the use of Binary Format for Scenes (BIFS) and by the Synthetic Natural Hybrid Coding (SNHC) subgroup for Animation Framework Extension (AFX). For example, AFX specifies the case of 3D virtual human animation with FBA and BBA (see Preda & Preteux, 2002).

Throughout this article, we will use the example of animated 3D virtual humans, for two reasons: First, it is a complete and consistent example of a 3D graphics application; second, because the presented concepts are easily extendable to other applications in 3D. One must note that while other standards, such as VRML and X3D, are more dedicated to 3D data than MPEG, our discussion lies in the context of MPEG because MPEG-4 can be considered as a superset of VRML. Overall, MPEG provides a complete framework for multimedia delivery, and recently, with the MPEG-7 and MPEG-21 standards, it even allows for the inclusion of media objects semantics as well as evolved processing, such as Digital Item Adaptation (DIA) or intellectual property management and protection (IPMP).

Moreover, adaptation is probably one of the most important issues considering 3D graphics delivery, and it offers a wide range of possibilities; for instance, by driving single content delivery towards multiple devices. The underlying approach uses a description of scalable encoded data to allow for bitstreams modifications (as illustrated by Figure 4), which is often done by using generic Bitstream Description Language (gBSDL), explained by Amielh and Devillers (2002). Though important work on spatial scalability has been done in Computer Vision by Lindeberg (1994), for instance, adaptation can be processed for many different types of media, as proposed by Gioia et al. (2004). One must focus on the particularity of individual media, such as audio as discussed by Aggarwal, Rose and Regunathan (2001), or video by Kim, Wang and Chang (2003), to ensure an optimum adaptation, both for the quality of content provided and for the accuracy of the adaptation. Recently, graphical adaptations have started to be designed by Raemdonck, Lafruit, Steffens, Otero-Perez and Bril (2002), and by Boier-Martin (2003), but many issues still remain, especially when considering the factor of real human perception, as described by Adelson (1991).

Furthermore, the large and growing variety of today's platform is a very important factor to consider when widely distributing 3D graphics. Some researches are oriented towards optimization for a specific device; other work is taking advantage of some computer graphics (CG) processing, such as Image-Based Rendering of Chang and Ger (2002), to allow 3D playback on light devices. Other approaches use these devices as a display only. For instance, Lamberti, Zunino, Sanna, Fiume and Maniezzo (2003) render 3D scenes with a cluster of machines and then transmit the rendered images to a PDA to be displayed. Recently, important work has been carried on for appropriate graphics API on light devices, the major one being probably OpenGL ES for embedded and mobile devices.

PRODUCTION OF 3D GRAPHICS

To provide immersive virtual experiences, 3D graphics contents must be carefully crafted and produced. Though a detailed description of the entire available production methods and pipeline would require many pages, such processes are required by every 3D-based applications, and thus are briefly discussed here.

The production of 3D graphics roughly consists of two main stages: the design of 3D shapes, and the creation of animation sequences applied on 3D models. Details on these stages follow, with a focus on the production of scalable contents to enable their adaptation in the delivery stage and an efficient playback for the client.

Representation

The standard techniques for designing 3D shapes, often referred as the global term of modeling, range from 3D scanning to user-designed models, with the help of 3D modeling software (e.g., 3DStudio Max, Maya, etc.). With today's technologies, it is also possible to create a virtual clone, using real pictures to create virtual humans according to anthropomorphic parameters and so forth, as shown in Figure 1. Scanners are usually used to produce a first-draft version of a mesh, which is then refined and com-

Figure 1. The left group of pictures illustrate the production of 3D models by cloning (real pictures on the left, extraction of texture and topology in the middle, virtual model on the right). The right group of pictures demonstrate the production of 3D models by anthropomorphic parameters.

Subject	(a)	(b)	(c)	(d)
Height	184	170	154	174
Neck	34	32	32	33
Bust	92	87	85	91
Underbust	76	75	71	74
Waist	75	68	73	63

pleted by artists or graphics designers in appropriate modeling software packages.

At this level, the mesh typically is represented using an ASCII or binary file in a proprietary format. For interoperability and consistency, VRML is a good choice. It often is selected as the 3D representation open format, especially in the context of MPEG, since VRML files are directly encodable into BIFS, and thus provide a good basic for adaptation. Furthermore, as illustrated in Figure 2, producing multiresolution meshes is a necessity to ensure adaptation of 3D shapes in the delivery stage.

Animation

To produce animation data, one can provide geometric input for all vertices of a 3D model, or position/orientation data to set rotations for joints and bones; for instance, in the case of body animation. Such methods consume a lot of time and human resources. On the other hand, motion capture systems (see Figure 3 for an optical system) output very realistic animation data, since the data are extracted from real motions in a more automatic way. This method is often preferred for creating

Figure 2. Examples of meshes at different resolutions. Top, from left: body models with 71K, 45K, 30K, 15K and 5K polygons. Bottom, from left: face models with 7K, 5K, 3K, 2K and 0.5K polygons.

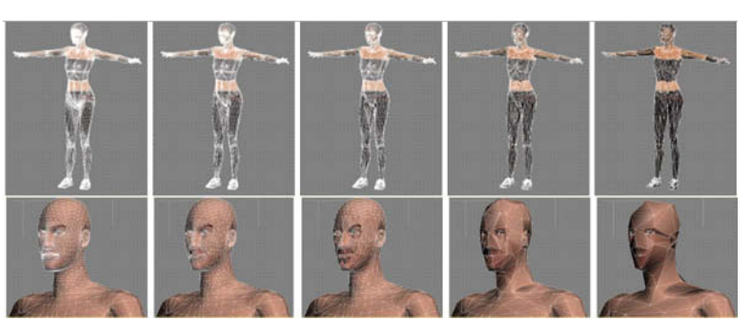
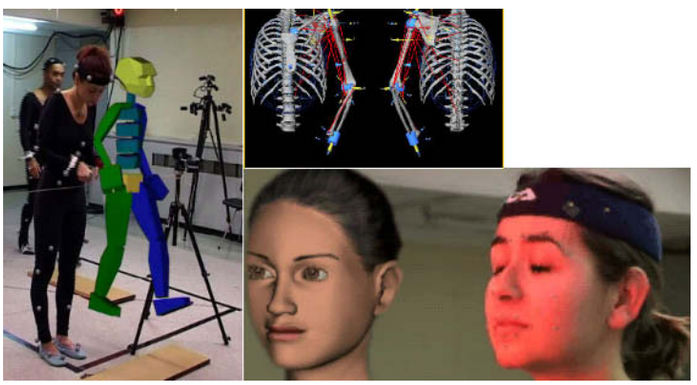


Figure 3. Skeleton-based animation and parameters from motion capture (for face and body)



animation, both for facial and body motions, though it requires manual fine tuning to meet the applications requirements, for instance, or to delete small artifacts that might arise from motion capture data (due to the noise, occlusions during captures, etc.).

DELIVERY OF 3D GRAPHICS

Similarly to audio, video and other media, networked architectures are well suited for the delivery of 3D content, should the data be streamed or downloaded.

Adaptation mechanisms provide a technological framework for efficient delivery of compressed 3D graphics and, thus, even in critical situations, such as lightweight devices and low bit rate, it allows for such a delivery. The adaptation of content for various networks and terminal capabilities as well as for different user preferences, is a key feature that needs to be considered and used as much as possible. Current state-of-the art research in adaptation shows promising results for specific purposes and types of content, and appears to be adaptable for massive heterogeneous environments. We propose here theoretical and practical methods for transmitting adapted 3D contents from shapes to animations to multiple target devices. The discussion is based on an integration of scientific methods into MPEG-21 and MPEG-4 architectures and an overview illustrated by Figure 7.

Figure 4. From left to right: Binary to bitstream description conversion; bitstream description to adapted binary data; a sample adaptation process at the bitstream level

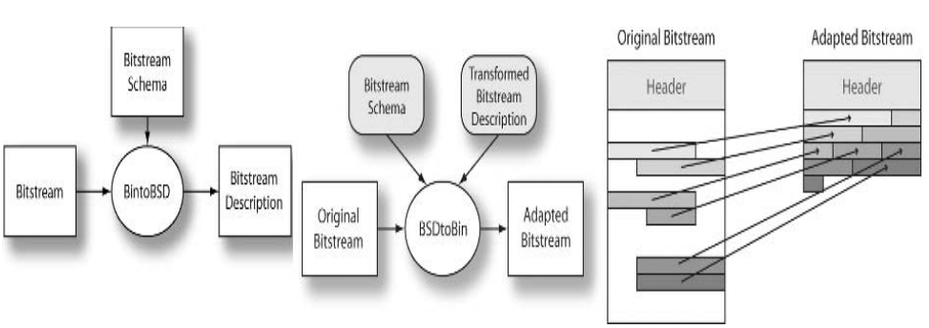
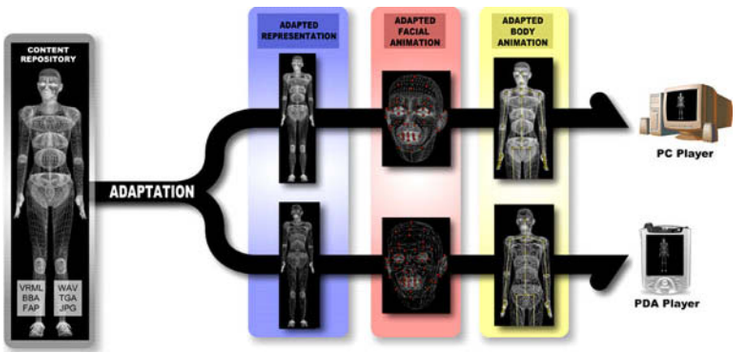


Figure 5. Adaptation process overview, featuring scalable 3D meshes as well as scalable facial and body animation data



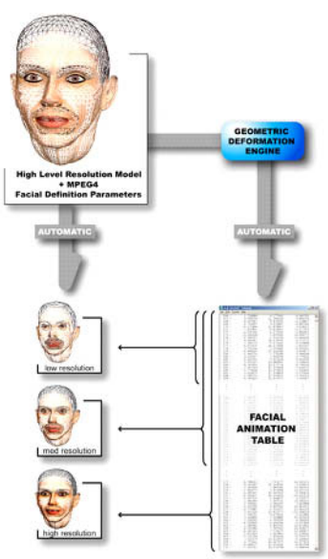
Adaptation Mechanisms

As illustrated by Figure 5, for a complete adaptation process, we must explore mechanisms for the following resources:

- 3D shape and representation; that is, 3D models and textures.
- 3D motion; that is, animation data for body and face when considering virtual humans.

One common factor in these two branches is the need for a very high-resolution base dataset. It will thus allow for the best possible quality on powerful machines while providing a suitable basis for simplifications.

Figure 6. Multi-resolution methods for scalable facial animation.



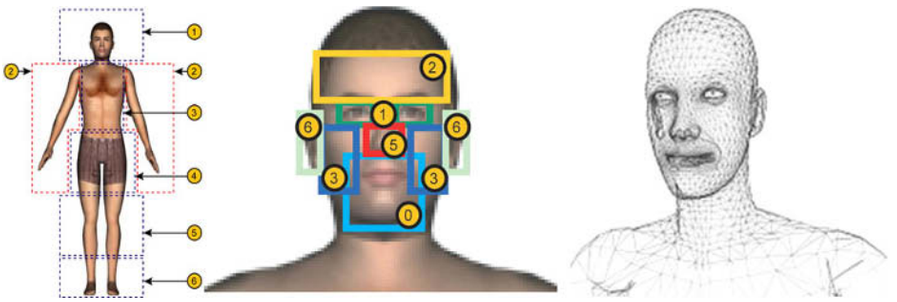
For the adaptation of 3D shapes, the methods used are based on a widely available and known technique in CG; namely, Level of Detail (LoD). Many different approaches have been proposed for LoD, such as the ones by Hoppe (1996), but without focusing on a use within adaptation frameworks. This issue has been explored by taking advantage of clustered, multi-resolution, encodable 3D models by Kim, Joslin, Di Giacomo, Garchery and Magnenat-Thalmann (2004).

For the adaptation of body animation (as well as facial animation), as illustrated by Figure 6, the refinement and simplification processes consist of the reduction or addition of joints; respectively, facial animation parameters. Such methods are detailed by Di Giacomo, Joslin, Garchery and Magnenat-Thalmann (2003), but basically consist of the removal of appropriate parts of the animation bitstreams and referenced as Level of Articulations.

Context Awareness

Within our scope, context is a generic word embedding a multitude of factors, including the constraints and/or benefits that drive the adaptation. Generally, it encompasses the network characteristics, such as the possible bandwidth and network restriction; and the terminal capabilities, such as its CPU frequency, its memory size and the user preferences and environment. Figure 7 illustrates the adaptation in interest zones that might be directed by the user preferences

Figure 7. From left: Interest zones for body animation; interest zones for facial animation; a high-resolution face on a lower resolution body due to the user's interest for facial animation and speech.

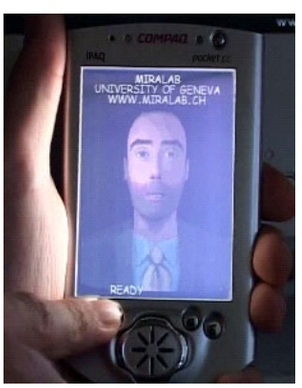


or focus (for instance, an interest for speech, and thus, facial animation at a particular time). There are many other different possibilities to use the context information, which has a set of criteria to drive the adaptation. One can, for instance, take advantage of a given bandwidth to dynamically reduce file sizes by removing high-resolution data for 3D meshes. One can also reduce the animation complexity to meet current CPU loads, and so forth. For more theoretical details on such issues, refer to Lerouge, Lambert and Van de Walle (2003), who present solving systems based on Pareto algorithms.

PLAYBACK OF 3D GRAPHICS

The production and delivery of graphics content might be rather independent from the target client's hardware, but the playback of such media is closely linked to the graphics capabilities of devices. See Figure 8 for a sample playback on a PDA. Not only is this stage traditionally hardware dependant, but it is also of paramount importance, due to the processing and memory resources required, and due to its intrinsic real-time constraints.

Figure 8. 3D rendering of a facial animation on PDA. In this example, the face model consists of a thousand polygons.



Assuming that the adaptation step, processed during the delivery, was driven by a context including graphics capabilities of the terminal, through some pre-benchmarking, the adaptation should ensure that the content delivered to the client is playable in real time.

Several different graphics API are available to develop a rendering engine for 3D players. Unfortunately, once again, these APIs are most of the time platform dependent, and though many are available on standard PCs (e.g., OpenGL or DirectX), mobile devices are not yet equipped with many options. OpenGL ES is probably the current most interesting work to provide a common and standard graphics API on lightweight devices. Furthermore, the group behind OpenGL ES is composed of many of the leaders in the 3D and graphics software and hardware industries.

CONCLUSION

Ensuring the delivery and playback of 3D graphics on the widest networked and terminal configurations open new and innovative roads to applications, not only in terms of pure 3D graphics such as games and virtual guides or assistants, but also in terms of the

Figure 9. Illustration of an AR cultural heritage application. Reality is extended by virtual characters and virtual life. The virtual objects are mixed with the real ones and displayed to the end users with Head-Mounted Displays.

combination of 3D with other media types. Sound is already commonly integrated in virtual scenes, but mixing video is a first step to allow for complex applications based on Augmented (or Mixed) Realities (AR or MR). For instance, as illustrated in Figure 9, ancient life can be added to real historical sites for education or tourism; industrial on-site training can benefit from virtual objects or instructors integrated in the real environment; and so forth.

ACKNOWLEDGEMENTS

Parts of this research and work have been funded through the European project DANAE (IST-1-507113) by the Swiss Federal Office for Education and Sciences (OFES). The authors would also like to thank Lionel Egger for his design and artistic talents.

REFERENCES

Adelson, E. (1991). Mechanisms for motion perception. *Optics and Photonics News,* (8), 24-30.

Aggarwal, A., Rose, K., & Regunathan, S. (2001). Compander domain approach to scalable AAC. *110th Audio Engineering Society Convention,* The Netherlands.

Amielh, M., & Devillers, S. (2002). Bitstream syntax description language: Application of XML-schema to multimedia content adaptation. *The 11th International World Wide Web Conference, USA.*

Boier-Martin, I. (2003). Adaptive graphics. *IEEE Computer Graphics & Applications*, 6-10.

Chang, C., & Ger, S. (2002). Enhancing 3D graphics on mobile devices by image-based rendering. *IEEE Third Pacific-Rim Conference on Multimedia, Taiwan.*

Di Giacomo, T., Joslin, C., Garchery, S., & Magnenat-Thalmann, N. (2003). Adaptation of facial and body animation for MPEG-based architectures. *IEEE International Conference on CyberWorld, Singapore,* 221-229.

Gioia, P., Cotarmanac'h, A., Kamyiab, K., Goulev, P., Mamdani, E., Wolf, I., Graffunder, A., Panis, G., Hutter, A., Difino, A., Negro, B., Kimiaei, M., Concolato, C., Dufourd, J., Di Giacomo, T., Joslin, C., & Magnenat-Thalmann, N. (2004). ISIS: Intelligent Scalability for Interoperable Services. *IEE 1st European Conference on Visual Media Production, England,* 295-304.

Hoppe, H. (1996). Progressive meshes. *SIGGRAPH, ACM SIGGRAOH, USA,* 99-108.

Joslin, C., Di Giacomo, T., & Magnenat-Thalmann, N. (2004). Collaborative virtual environments, from birth to standardization. *IEEE Communications Magazine, Special Issue on Networked Virtual Environments, 42*(4), 65-74.

Kim, H., Joslin, C., Di Giacomo, T., Garchery, S., & Magnenat-Thalmann, N. (2004). Multi-resolution meshes for multiple target, single content adaptation within the MPEG-21 framework. *IEEE International Conference on Multimedia & Expo, Taiwan.*

Kim, J., Wang, Y., & Chang, S. (2003). Content-adaptative utility based video adaptation. *IEEE International Conference on Multimedia & Expo., USA.*

Lamberti, F., Zunino, C., Sanna, A., Fiume, A., & Maniezzo, M. (2003). An accelerated remote graph-

ics architecture for PDAs. *Web3D 2003 Symposium, France,* 55-62.

Lau, R., Li, F., Kunii, T., Guo, B., Zhang, B., Magnenat-Thalmann, N., Kshirsagar, S., Thalmann, D., & Gutierrez, M. (2003). Emerging Web graphics standards and technologies. *IEEE Computer Graphics & Applications*, 66-75.

Lerouge, S., Lambert, P., & Van de Walle, R. (2003). Multi-criteria optimization for scalable bitstreams. *8th International Workshop on Visual Content Processing and Representation, Spain*, 122-130.

Lindeberg, T. (1994). *Scale-space theory in computer vision.* Dordrecht, The Netherlands: Kluwer Academics Publisher.

Preda, M., & Preteux, F. (2002). Advanced animation framework for virtual character within the MPEG-4 standard. *IEEE International Conference on Image Processing, USA,* 509-512.

Van Raemdonck, W., Lafruit, G., Steffens, E., Otero-Perez, C., & Bril, R. (2002). Scalable 3D graphics processing in consumer terminals. *IEEE International Conference on Multimedia & Expo., Switzerland,* 369-372.

Walsh, A., & Bourges-Sévenier, M. (2001). *MPEG-4 Jump-Start.* USA: Prentice Hall.

KEY TERMS

AFX: Animation Framework eXtension. Standard conducted by the Synthetic and Natural Hybrid Coding group within MPEG, its goal is to specify compression schemes for 3D animation data and tools, such as body animation, image-based rendering, texture-mapping and so forth.

BBA: Bone-Based Animation. This AFX tool allows for the compression of virtual human animation, including skin deformations, and also for generic hierarchical animation.

BIFS: Binary Format for Scenes. Based on VRML97, BIFS is extended with commands that can update, delete or replace objects in the scene. For streaming scenarios, BIFS also offers an integrated binary compression scheme, media mixing and audio composition.

FBA: Facial and Body Animation. This MPEG-4 system tool was designed for the compression and networked delivery of virtual human animation. The facial animation part is still in use, while the body animation part is being replaced by BBA for sake of genericity.

MPEG: Moving Picture Experts Group. MPEG is a working group of ISO/IEC in charge of the development of standards for coded representation of digital audio and video. Since 1988, the group has designed MPEG-1 (used for video CDs and MP3s), MPEG-2 (used in digital television set top boxes and DVDs), MPEG-4 (used for mobile use cases and networking), MPEG-7 (for the description of digital content) and MPEG-21 (which mainly addresses rights management, digital items and adaptation of content).

OpenGL ES: OpenGL for Embedded Systems. OpenGL ES is a low-level, lightweight API for *advanced embedded graphics* using well-defined subset profiles of OpenGL. It provides a low-level applications programming interface (API) between software applications and hardware or software graphics engines.

VRML: Virtual Reality Modeling Language. VRML is an open 3D description language, whose first version became an international standard in 1995. Developed by the Web3D consortium, its goal is to allow shared virtual worlds and 3D media on the Web.

X3D: eXtensible 3D graphics. X3D is an open standard for Web 3D-Graphics, whose primary goal is to express the geometry and behavior capabilities of VRML, using XML for description and encoding.

Public Opinion and the Internet

Peter Murphy
Victoria University of Wellington, New Zealand

INTRODUCTION

The development of the "World Wide Web" has had a significant impact on the formation of public opinion in democratic societies. This impact, though, has not been exactly that predicted by early 1990s prophets of the Web, who expected a decentralization of traditional mass media. If anything, the easy accessibility of the Web-enabled Internet (hereafter "the Net") has extended the audience reach of traditional network media. Despite this, the Net is fundamentally changing the nature of public opinion.

One should be wary of thinking of this change as a technology-enabled extension of the 19th-century liberal public. In the liberal view, the Net is a difficult-to-control free speech medium. It engenders a babble of voices devoted to persuading citizens and governments of the merits and otherwise of laws and policies. Because the Web's infrastructure of servers is global, dictatorial, or even legal control of it is difficult to achieve. This is especially true for governments that want to encourage the pragmatic benefits of computer-mediated commerce.

Yet, to see the Net simply as a free speech medium does not do full justice to its nature. It began life as a powerful document delivery system, and, in important ways, its long-term impact on public opinion derives from that fact. The Web leveraged existing inter-networked computing to enable a new way of creating, collecting, storing, transforming, and disseminating documents and information objects. The frothy activity of instant commentary and interest group campaigning that the Net facilitates disguises the extent to which the logic of the public sphere is undergoing a long-term paradigmatic shift shaped by its origins as a document archive.

BACKGROUND

The architect of this dynamic document archive was Tim Berners-Lee (Berners-Lee, 1999; Naughton, 1999). In 1980, Berners-Lee began work as consultant at CERN, the international particle research body located near Geneva. CERN was a "city of turnover". Its principal social characteristic was a transient population. Visiting physicists who came and went did much of the center's experimentation. Scientists on average stayed two years. The problem that resulted was how to maintain good documentation tracking when staff turnover was so high. Berners-Lee set out to solve this problem.

His first attempt was a program called ENQUIRE (1980), which he called a "memory substitute". He filled documents with words which, when highlighted, would lead to other documents. This was similar to the Apple Macintosh HyperCard. This application in its turn borrowed the hypertext concept from Ted Nelson (Nelson, 1992). Hypertext conceived information as connection or linkage. Berners-Lee adapted this idea to create the beginnings of a publicly accessible archive of documents. The archive was initially restricted to CERN. In 1989, however, Berners-Lee conceived a plan for a universal document system. Universal meant global. The idea was to use a mix of hypertext and networked computing to link all documents and information objects in the world. The idea of a universal system was a conceptual break-through. A universal system meant there would be no central control or source of information—whether in the sense of ea centralized undemocratic hierarchy or else a democratic hub-and-spoke network. Universal also meant the potential integration of all information systems.

Berners-Lee had another powerful idea. He thought that a universal information system should mean not only universal access to and retrieval of documents but also the universal capacity to publish documents. He insisted (against the opposition of peers) that this should be a system in which anyone using a hypertext editor could publish a linked document. The hypertext editor was the forerunner of the HTML editor. Andries van Dam had created the first functional hypertext editor in 1967 at Brown University.

In 1990, Berners-Lee got support from CERN senior managers for what had been to that date

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

virtually a private project. He created a program called a "browser" that provided a virtual "window" through which a user saw a Web of linked resources on the existing "inter-net" (i.e., the existing internetwork of networked computers that had grown up since the 1970s). His small team also created a "Web server", based on the client-server model. He envisaged a system in which information would be stored on networked computers called servers. Client programs (browsers) running on other networked computers would access these servers.

How would the information be extracted from these servers? One option was to use existing technology such as TELNET or FTP. A second more powerful idea was that of the "inter-face". This concept came from the hypertext community. An inter-face was a "window" that displayed the structure of the virtual space of linked texts. Originally, node-link diagrams represented this structure. The first browsers were not graphical. Graphical interfaces came later. Marc Andressen's 1993 Mosaic browser was the first with the standard graphical interface of windows, graphics, and point and click functionality.

Berner-Lee's desire for universality meant that he had to ensure that public information on any networked computer anywhere in the world could be accessed through the browser. To achieve this end, Berners-Lee devised a set of protocols by which different machines could talk to each other and exchange information. One protocol specified the location of information. It was like an IP address. A second protocol for information exchange between machines was modeled on FTP. This was the HTTP (Hypertext Transport Protocol). A third protocol established a uniform way of structuring documents: Hypertext Mark-up Language (HTML). HTML was based on SGML (Standard Generalized Mark-up Language) already used in the electronic publishing world. It provided conventions for attaching tags to pages.

CRITICAL ISSUES: FROM PEERS TO AUTOPOIESIS

The result of Berners-Lee's architecture was a cheap, quick, and reliable system for accessing, retrieving, and publishing documents. Any person with access to the Internet in principle could look at any document stored on a Web server (unless it resided on a secure server where access was intentionally limited). A person with some Web server space could publish any documents they liked on the Internet, as long as they had some simple knowledge of HTML page creation.

What followed from this were two major consequences for public opinion. The first was that anyone with a relatively simple set of tools could publish their own opinions. On the Web, these opinions were accessible to anyone anywhere in the world with access to a computer and an Internet Service Provider.

Computer-mediated universal access and self-publishing created a new kind of public sphere. They also created a new set of justice and equity problems. Not everyone can afford access, and certainly not unlimited access, to the Internet. Indeed most of the world does not have a telephone connection, let alone a computer or an ISP account. But, then, also most of the world has never participated in public opinion formation of any kind. In the still limited number of countries where there is a history of strong public spheres, programs sponsored by governments and private foundations emerged in order to overcome access inequalities. Widespread provision of computing by companies and educational institutions also facilitated access to the new digital public as well. "Stealing time" from institutions for public and private Net activity emerged in the well-endowed democracies as a "quasi entitlement"—creating dilemmas for organizations as to "when and where and how" to encourage or discourage such tacit activity.

In democratic societies with long-established publics, and a correlative strong propensity to create intellectual wealth, virtually all social groups and classes have directly or indirectly benefited from the increasing access to information made possible by the Net. At the other end of the political spectrum, the Net has posed significant dilemmas for dictatorial governments. Their first instinct has been to censor Web materials. However, censorship is difficult to apply to the Net, because material is published on thousands of web servers in hundreds of countries. Dictatorial states instead discourage access to computer hardware, the setting up of ISPs, and the local publication of sites. However, as the Net is also a major scientific and commercial medium, with implications for trade and military science, such controls also hurt a state's economic and technology performance.

In contrast to crude dictatorships, authoritarian states like China have sought to preserve the economic and scientific advantages of the Net, while discouraging free speech and restricting freedom of information. Such states encourage user accounts while maintaining a state monopoly over government documents, blocking access to a relative handful of politically sensitive international news and government sites, and closing down local opposition sites. These measures alone cannot prevent individuals browsing critical materials, so authoritarian states have come to rely heavily on the strategy of self-censorship. They rely on the fear of Web users that the government may find out about, and punish them, for visiting sites that the government disapproves of. Users are aware that it is difficult to erase all traces of such activity from a computer's hard drive. Packet sniffing, keystroke monitoring, and inspection of logs allow systems administrators to audit unauthorized activity on network computers. But monitoring all Net activity would be insanely labor intensive, and thus self-defeating for any government. So authoritarian states depend on their population using the Net for social communication (for chat) but not for political communication. A government might occasionally audit the immensely popular chat rooms, but so long as users avoid explicit political comment the state has no further interest in what is being said.

The success of the strategy of self-censorship has been one of the reasons that the Net has not proved to be the kind of libertarian force that its prophets in the mid-1990s expected it to be. However, authoritarian state strategies are not the only reason for this. A lot of the Net's supposed power to shape public opinion is overstated. Take the much-touted ability of Net users to post opinions on the Web. Anyone in a democratic state with modest resources and motivation can publish more or less what ever they like, more or less where-ever they are, and at any time. The popularity of Web logs, threaded discussions, relay chat, and so on are testament to this. However, often this means little more in practice than that the Net is a powerful expressive medium. It allows no-holds-barred statements of opinions and views. Other Net users, though, can just as easily ignore these. Cohorts involved in threaded discussions typically have difficulty sustaining dialogues. It is striking how minimally interactive much supposed interactive discussion actually is (Davies, 2003: 37-38). Expression on the Net is often mistaken for discussion. This phenomenon has significant implications for the Net as a medium of public opinion.

Net citizens, or netizens (Hauben, 1997), have difficulty sustaining arguments with political opponents. They quickly drift off topic. They don't engage with each other. History can help us understand why this is so. Peer-based formation of public opinion emerged in the 18th century (Habermas, 1991). It arose out of face-to-face debates that had been released from the constraints of traditional social hierarchies. Coffee houses and the houses of parliament in London were crucial to this development. In this setting, we see public opinion formed through the arguments of peers. Peers have no social authority to compel others to agree with their opinions. As in a jury room, they have to garner agreement by reasoning. In 18th-century England, newspapers recorded the debates of peers. Thanks to existence of an effective postal service, the reports of these debates could be sent to the provinces. Debate between peers meant that public opinion was shaped by feedback. Statements were made, and others responded to them. Responses in turn were responded to, as the pitch of debate increased.

It is an illusion to think that the Net functions like this. It has many powerful tools to facilitate interactive responses—from discussion boards to e-mail. But the ability of these tools to reach anyone with an e-mail address also means that the technology contradicts the small-scale logic of peer debates. The greatest extent of one's peers is around 150-200 people. Yet the Net allows everyone in the world to be one's peer. Peer-style feedback cannot function meaningfully on that scale. Cybernetic models of feedback may work for machine self-regulation (Weiner, 1948) but not for opinion articulation.

The world scale of the Net is the result of a longer historical process in which the small scale of peer debate has been subsumed by larger-scale processes of public opinion formation. From the mid-19th century, telegraphic (and later telephony) networks permitted news services to transmit opinion samples to news organizations with great speed and "from anywhere to anywhere" served by these networks. Correspondingly, newspapers developed editorial formula to communicate with a mass audience, in place of peer audiences. The rise of the organization society and its generic ideologies—such as liberalism

and socialism—abetted this development. Communication became a professional activity. With the development of radio and television networks, formula-driven reports could be instantly transmitted to a mass audience. The public opinion that developed in this context was formed through the gatekeeping of competing news organizations. How opinions were collected, edited, and redistributed through networks of public broadcasters and private media companies was crucial to their eventual shape.

The third, most recent, stage of public opinion emerged with the Web in the 1990s. Gatekeeper publishing organizations have a strong presence on the Net. Peer-to-peer forums and tools are also widely available and well supported. However, the key innovation of the Net is that virtually anyone with a basic skill set and modest resources can publish their own material. They can "post" material to a URL (Universal Resource Location) address. Each byte of data in a computer memory has a numeric address. Addresses allow data to be located. The model of the Net as an addressable medium was initially derived from Von Neumann's computer architecture (Bolter, 1984; Floridi, 1999). The idea of the numeric addressing of space ultimately derives from Descartes. Long before computing, it underpinned the modern concept of a postal service with its numeric street addresses and zip codes. When Berners-Lee adapted this "reading, writing, addressing, and posting" technology, what we ended up with was individuals being able to "post" a document to a public computer address that anyone could browse. As long as a person was motivated to search for the document that might be located at any of millions of addresses, and as long they had some search skills and tools, they could locate the document.

What the "public post" model is geared to is not peer-to-peer communication but archival transmission. It is not governed by the judgment of professional editorial gatekeepers but by self-publishers. This begs the question: how is public opinion formed in the age of addressable media?

One answer to this question is that addressable media do not support the type of collective public opinion typical of the age of large media organizations. Partly this is because of the reduction in the influence of editorial filtering mechanisms that can shape such an opinion. Partly this is because collective opinion is simply less important to democratic functioning in a cybernetic society in contrast to the growth of self-regulating systems. One of the most important examples of a self-regulated system is the Net itself. What counts is not its capacity for broadcasting opinion, or for stimulating mutual dialogue. What is crucial is its capacity to post, archive, and retrieve opinion in a self-regulating way. The Net makes us rethink the very nature of opinion.

The Net is a self-organizing or autopoietic system. The classic example of such a system is the city (Jacobs, 1985; Johnson, 2001; Murphy, 2001, 2003). For example, the way that traffic flows in a city exists independently not only of each driver's desires but even of the intentions of the most foresighted planner. Little that happens in cities is explicitly legislated, yet city life is shaped by powerful patterns well understood by its denizens. Symmetry, proportionality, and economy generate many of these anonymous forms. They can last for generations. Some of the patterns of Rome, for example, have persisted for over 2,000 years. Such patterns often prove highly palatable to city dwellers. They make good use of them to generate their own incremental additions to city life.

The Net operates much like a city. We can begin to understand why this is so if we look again at Berners-Lee's original design for the Web. He designed it to archive documents. Its purpose was to transmit science documents over time. Scientists who left CERN could archive their papers so that they would be readily available to incoming researchers. The model of this was neither the debating forum of scientific peers, nor was it the office newsletter. What emerged from the initial design of the web was a giant Alexandrine-like archive. The things that characterize the archive are:

1. It is driven by the self-publishing and self-organizing efforts of its contributing parts. No contributing part (individual, group, or organization) has much influence measured against the whole of the Net. No contributing part can be a gatekeeper for the whole. There is not an editorial "ghost in the machine" to regulate the system. Likewise, the archive has no peer bodies (for example, a Senate or Dr. Johnson-style clubs) where public opinion is decisively shaped.
2. In self-organizing systems what counts is the long-term transmission of pattern and structure. Generations come and go, endless changes are

made, and yet through all of the changes certain patterns persist. The contribution of each part belongs to a larger scheme of things.

3. Each part has difficulty comprehending the whole of the archive, but each contributor nonetheless still understands something of its tacit architecture.
4. This architecture, like all great architecture, is simple. With a few elementary pattern-ideas, beginning in the case of the Net with a few protocols, a complex structure is created.
5. Other patterns emerge spontaneously—like the Zipf distributions or "power law" of the Net.
6. Like a city, sight and sound and movement are as important to the Net as text is. Correspondingly, opinion that lasts is as much characterized by its composition and design as by its peer standpoint or its generic ideology.
7. Such an autopoietic system allows millions of persons to contribute to it. The nature and meaning of the system remains independent of the intentions, beliefs or opinions of any and all of the contributors. Like a city, the autopoietic archive has a character separate from its makers.

CONCLUSION

Peer opinions emerged as important in the collegial societies of the 18th century. Editorial opinion became crucial in the context of the organizational societies of the late 19th and 20th centuries. As it entered the era of the archive, opinion assumed the time-scale of autopoietic systems. This is "the thousand-year scale of the metropolis" (Johnson, 2001, p. 99). The future lasts a long time. The power of such an archive is transmissive rather communicative (Debray). What matters is not the communicative interaction of peers, or mass communication, but transmission across time.

Transmissive power is measured in decades, centuries, and even millennia. The medium of the Net has exceptional capacity to instantly send, retrieve, and self-publish material. Yet the ultimate logic of the Net is to preserve and transmit those documents and objects over time. An understanding of large-scale transmissive systems is still sketchy. It remains a key topic for future researchers.

REFERENCES

Berners-Lee, T. (1999). *Weaving the Web.* New York: HarperCollins.

Bertalanffy, L. (1968). *General systems theory.* New York: George Braziller.

Bolter, D (1984). *Turing's man.* Chapel Hill: University of North Carolina Press.

Davies, W. (2003). You Don't Know Me, but … Social Capital and Social Software London: Work Foundation.

Debray, R. (2000 [1997]). *Transmitting culture.* New York: Columbia University Press.

Habermas, J. (1991). *The structural transformation of the public sphere.* Cambridge, MA: The MIT Press.

Jacobs, J. (1985). *Cities and the wealth of nations.* New York: Vintage.

Johnson, S. (2001). *Emergence*. Harmondsworth: Penguin.

Hauben, M. & R. (1997). *Netizens*. Los Alamitos, CA: IEEE Computer Society Press.

Murphy, P. (2001). *Civic justice.* Amherst, NY: Humanity Books.

Murphy, P. (2003). Trust, Rationality and the Virtual Team. In D. Pauleen (Ed.), *Virtual teams: Projects, protocols and processes.* Hershey, PA: Idea Group.

Naughton, J. (1999). *A brief history of the future.* London: Weidenfeld & Nicolson.

Nelson, T. (1992 [1981]). *Literary machines 93.1.* Watertown, MA: Eastgate Systems.

Weiner, N. (1948). *Cybernetics*. New York: John Wiley.

KEY TERMS

Autopoietic System: A self-making or self-organizing system.

CERN: The High Energy Physics Laboratory, Geneva.

Dialogue: A conversation and exchange of opinion between two or more persons.

FTP: File Transfer Protocol, a convention for the transfer of files across computer networks.

Peers: Persons who are equal in social or occupational standing.

Power Law: Postulates that Internet goods (income, traffic, links, etc.) will follow Zipf-type distributions. For example, this predicts that a second-ranked site will be half as successful as a first-ranked site, and a 10th-ranked site will be a 10th as successful as the first-ranked.

TELNET: An early protocol for network computing, enabling a user to logon to a remote machine and access its files.

Search Engine: A program that searches the web indexing and cataloguing the pages that it finds.

System: A structured activity in which the "whole" is more than the sum of its "parts".

Quality of Service Issues Associated with Internet Protocols

Jairo A. Gutiérrez
The University of Auckland, New Zealand

Wayne Ting
The University of Auckland, New Zealand

INTRODUCTION

The objective of enabling the development of higher-level multimedia services with guaranteed quality of service (QoS) on networks has prompted developments that attempt to accommodate these new application requirements. Several architectures have been proposed, and a common basic functionality is emerging. Any new architecture that intends to satisfy the ever-growing need for bandwidth in the Internet while providing support for QoS guarantees needs to concern itself with the following aspects (Zhang, Deering, Estrin, Shenker, Zappala, 1993; Biswas, Lazar, Huard, Lim, Mahjoub, Pau, Suzuki, Torstensson, Wang and Weinstein, 1998):

- Flow management: identifying the traffic characteristics of a flow so that the network can specify the QoS to be delivered to that flow
- Compatibility with a wide range of routing protocols (Callon, Doolan, Feldman, Fredette, Swallow, Viswanathan, 1997)
- Resource reservation
- Admission control
- Packet scheduling: including packet filtering and classification.

These aspects clearly call for network devices (routers, switches) with powerful features that can be easily requested and modified in order to support customers' demands for differentiated services.

BACKGROUND

The Internet phenomenon has grown at an exponential rate, causing several new technologies to emerge to cope with the number of users of this global communications network. TCP/IP was the protocol of choice because of its simplicity and ability to a route from a particular source to a particular destination. However, as the network grows, there have been greater demands for additional services and real-time applications to work over the Internet. Because TCP/IP is primarily a best-effort protocol, it is not able to provide the QoS required by real-time applications and its users. To complement this deficiency, other protocols, such as Integrated Services (IntServ), Resource Reservation Protocol (RSVP), Differentiated Services (DiffServ) and Multi-Protocol Label Switching (MPLS), have been developed.

IMPROVING INTERNET PROTOCOLS

The Integrated Services (IntServ) model is based on reservations-based traffic engineering assumptions. It reserves resources explicitly using a dynamic signalling protocol (RSVP) and employs admission control, packet classification and intelligent scheduling to achieve a desired QoS. The Integrated Services model has two services categories: Guaranteed Delay and Controlled Load services (Braden, Clark & Shenker, 1994).

RSVP is a resource reservation set-up protocol designed for an integrated service Internet. It is used by a host to request specific qualities of service from the network for particular application data streams or flows. RSVP is also used to establish and maintain state information in all nodes along a flow so as to provide the requested service. However, RSVP itself is not a routing protocol; instead, it is considered a signalling protocol similar to those used in ATM networks.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

The DiffServ model is based on reservation-less traffic engineering assumptions. It classifies packets into a small number of service types and uses priority mechanisms to provide adequate QoS to the traffic. No explicit resource reservation or admission control is employed, although network nodes do have to use intelligent queuing mechanisms to differentiate traffic (Gozdecki, Jajszcyk& Stankiewicz, 2003).

DiffServ allows levels of service discrimination but without the need to maintain per-flow states and signalling (Fineberg, Chen & Xiao, 2002). Packets are classified according to Behaviour Aggregates (BA), and these BAs are encoded using Differentiated Service's Code Points (DSCP). There is a one-to-one correspondence between a BA and a so-called Per-Hop Behaviour (PHB), which will define the actual treatment for that particular BA.

Another requirement for DiffServ is the identification of a well-defined boundary called DiffServ domain. The reason for this is so only the boundary nodes in the DiffServ domain need to classify packets into a particular BA, and maybe condition or shape the ingress traffic accordingly (Gozdecki, Jajszcyk & Stankiewicz, 2003). The result is that only the boundary node needs to carry out sophisticated classifying, marking, policing and shaping operations, while the core nodes only need to match the DSCP in the packet to a PHB and forward accordingly.

Consequently, the DiffServ approach produces a more scalable architecture by having groups of packets with similar QoS requirements matching with a single BA; hence, the number of states kept is proportional to the number of classes defined rather than being proportional to the number of flows. DiffServ is also easier to implement because sophisticated operations are only carried out at the boundary nodes, and the core nodes only carry out the simple operation of matching DSCP to a PHB; therefore, the speed in the core can also be faster.

MPLS is a protocol that assigns a particular FEC (Forwarding Equivalence Class) to a particular packet as it enters the network. The FEC to which the packet is assigned is encoded as a short, fixed-length value known as a "label." Once a packet is assigned to a FEC, no further header analysis is done by subsequent MPLS capable routers and all forwarding is driven by the labels (Rosen, Viswanathan & Callon, 2001).

MPLS, used with or without RSVP (Zhang, Deering, Estrin, Shenker & Zappala, 1993; Baker, Krawczyk, Sastry, 1998), fits within the framework of Integrated Services (Braden, Clark & Shenker, 1994). Before suggesting detailed interfaces for IP routers or switches, it is necessary to understand the basic operations of this proposed standard. The roots of MPLS originated from technologies such as IP switching, developed by Ipsilon Networks (Sunnyvale, Calif.), and tag switching, developed by Cisco Systems (San Jose, Calif.). Thus, MPLS is the use of fixed-length labels to decide packet handling (Xiao & Ni, 1999). An MPLS router, called the label-switch router (LSR), examines only the label in forwarding the packet. The network protocol used can be IP or others. MPLS also needs a protocol to distribute labels to set up label-switched paths (LSPs). The protocol used to distribute labels is known as the label distribution protocol (LDP). Whether a generic LDP should be created or RSVP should be extended for this purpose is an issue yet to be decided (Xiao & Ni, 1999). MPLS can also be piggybacked by routing protocols. A LSP is similar to an ATM virtual circuit (VC) and is unidirectional from the sender to the receiver. MPLS LSRs use the protocol to negotiate the semantics of how each packet with a particular label from its peer is to be handled. Therefore, when a packet enters an MPLS domain, it is classified and routed at the ingress LSR. MPLS headers are then inserted into the packet. When an LSR receives a labelled packet, it will use the label as the index to look up the forwarding table. This is faster than the processes of parsing the routing table in search of the longest match done in IP routing. The packet is processed as specified by the forwarding table entry. The outgoing label then replaces the incoming label, and the packet is switched to the next LSR. Inside an MPLS domain, packet forwarding, classification and QoS service is determined by the labels and the COS fields. This makes core LSRs simple. Before a packet leaves a MPLS domain, its MPLS label is removed (Xiao & Ni, 1999). As far as the original packet is concerned, the routers carrying it through the MPLS network appear as a single hop (Stephenson, 1998).

Despite its advantages, MPLS does, however, have one major drawback as the protocol to implement QoS in the Internet. The architecture and protocols defined by MPLS require a much more

extensive change to conventional IP networks than the other protocols discussed so far. This need for major upgrades will certainly slow down the widespread implementation of MPLS, and will perhaps limit the use of LSRs to core routers in the Internet.

Because MPLS decouples the label distribution mechanisms from the data flows, it supports several physical and link-layer technologies. For ATM, the label is placed in the ATM cell header in the VPI/VCI field. In a local-area network (LAN), the label is placed after the MAC header. Standard routers can act as core LSRs by running Label Switching software (adding the capability to switch labelled packets based on the label values), and by supporting LDP.

AN INTEGRATED APPROACH TO END-TO-END QOS ON THE INTERNET

The Internet QoS protocols (Integrated Services, Differentiated Services, RSVP and MPLS) use different technologies to achieve their goal. Whether the protocols become widely adopted by hardware and software vendors largely depends on several factors. Some of the most important are the scalability of the protocol as well as the granularity of the service classes offered. (More on this in the Integrating MPLS and Differentiated Services section.)

In general, Integrated Services (Braden., Clark & Shenker, 1994) is more appropriate for networks with fewer flows, such as those found in LANs and corporate networks. These networks are usually considered to be at the boundaries of the Internet. Differentiated Services (Yoram, Binder, Blake, Carlson, Carpenter, Srinivasan, Davies, Ohlman, Verma, Wang & Weiss, 1999), on the other hand, is more appropriate for networks found further within the Internet. For example, ISP networks should implement Differentiated Services because the amount of flows that they would have to manage would be significantly larger than those found in corporate networks. Also, by the very nature of the protocol, ISPs would be able to provide differentiated services to their clients while still able to manage QoS efficiently. MPLS, because of its architecture, is more appropriate for routers and networks deep within the Internet, where the number of flows to handle is much greater than those found in Differentiated Services networks. With such great number of flows to manage, the fine granularity of service classes as well as the label switching technology will significantly improve QoS.

As for RSVP, it is the signalling protocol that has been mentioned by each protocol's working group to be a possible candidate for their signalling needs. It is already known that RSVP works very closely with the Integrated Services protocol to deliver end-to-end QoS. In fact, it is considered part of the Integrated Services architecture. It has been suggested that RSVP be adopted in Differentiated Service networks (Yoram et al., 1999) so as to provide end-to-end QoS over a Differentiated Service domain. RSVP has also been suggested as a candidate to be used with MPLS (Guerin, Gan, Kamat, Li & Rosen, 1997; Black, 2002) to set up Label Switched Path tunnels.

Integrating MPLS and Differentiated Services

The main problem for MPLS to support DiffServ is to decide how MPLS will map the BA and their unique DSCP to the labels in MPLS so that the LSR can determine which PHB to apply to the packet. To understand the various ways that the mapping is done, a few new concepts must be introduced to help our understanding. The first concept is that of an Ordered Aggregate (OA). An OA is defined in RFC 3260 (Grossman, 2002) as "The set of Behaviour Aggregates that share an ordering constraint"; that is, those BAs share the same scheduling characteristics. The second concept is a PHB Scheduling Class (PSC); this is again defined in RFC 3260 as "The set of one or more PHB(s) that are applied to the Behaviour Aggregate(s) belonging to a given OA"; that is, one OA will have one PSC that corresponds with all the PHB(s) in that OA. In RFC 3270, two ways of mapping DSCP to labels have been proposed. The first is to use the three bits in the EXP field of the MPLS header to map to the DSCP, and the actual label (which represents the FEC and the LSP that these packets will be travelling on) remains unchanged. The result of this is that a given LSP, can support up to eight Bas for a single FEC (Black, 2002). These LSPs are called "EXP-inferred-PSC LSPs" (E-LSP). For an E-LSP, the actual label will be used by the LSR for forwarding, while the EXP field will be used to determine which

PHB to apply to the packet. The other method proposed by RFC 3270 is to establish a separate LSP/label for a single [FEC, OA] pair and this type of LSP is called "Label-only-inferred-PSC LSP" (L-LSP). For L-LSP, the PSC is implied within the label, so the LSR can infer both the forwarding and scheduling information from the label and you are no longer restricted to the eight BA per FEC constraint (Black, 2002). L-LSP now uses the EXP field to carry the drop precedence if the "shim" header is used for MPLS; or, if the label values are contained in a layer-2 header, then the specific drop precedence field for that particular layer-2 technology will be set accordingly[1] (Black, 2002).

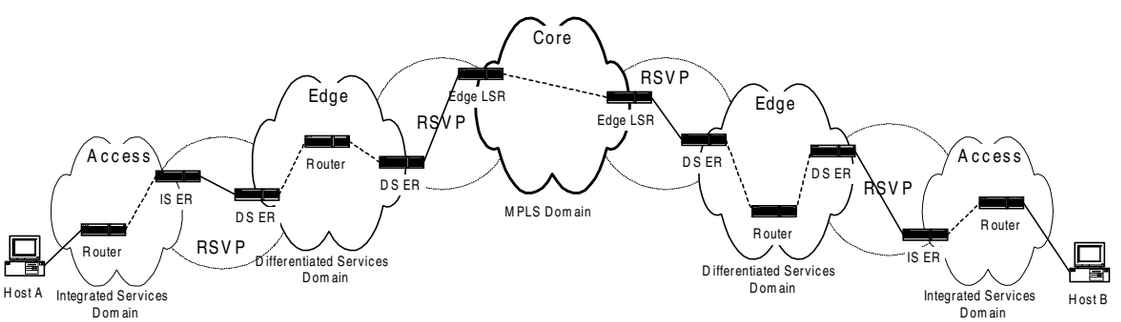
After describing how DiffServ can be mapped to MPLS, a walkthrough of how a packet will pass through a DiffServ MPLS domain is presented to show how everything is integrated. When a packet comes into an ingress DS-LER[2], the packet will first pass through a Classifier to determine first which BA/OA should be applied to this packet[3], and second, which FEC should this packet belong to. After the classifier, the packet may need to go to a Meter that will determine if this packet violates the Service Level Agreement (SLA) signed between the service provider and the customer. Then we arrive at the Marker, who will take information from both the Classifier and Meter to decide which particular MPLS label it will give to the packet according to its FEC, BA and SLA conformance. After that, it will go to the Shaper/Dropper, who will make sure the packet goes to the appropriate queues or potentially drops the packet if it does not conform to the SLA. Finally, the packet leaves to the Label Edge Router (LER) and into the core MPLS network. In the mean time, the signalling protocol for MPLS[4] is busy exchanging label information between LSRs so that a particular LSP can be set up (if one LSP for the FEC and with the QoS requirements are not already there) for our packet to travel on. After leaving the ingress LER, the packet travels through various core LSRs who carry out label swapping to forward the packet towards its destination egress LER. When the packet arrives at the egress LER, all MPLS labels and headers are removed at the egress node and the packet is delivered out of the MPLS domain to the destination.

AN INTEGRATED APPROACH TO QOS

TCP/IP, though lacking in QoS characteristics, is required as the common protocol used by the approaches discussed previously. The QoS features of ATM, on the other hand, give us an idea of the characteristics that each of these approaches is trying to achieve. With this in mind, the following diagram illustrates a framework of protocols and their possible mapping with each other to provide end-to-end QoS.

From Figure 1, a data packet originating from an Integrated Services domain, requiring a certain level of QoS, may traverse through several other networks before reaching its destination. RSVP may be included if the flow requirements include the reserva-

Figure 1. An integrated approach to QoS



tion of resources. These networks might be Differentiated Services networks or MPLS networks. However, if there is a service mapping between the different networks, end-to-end QoS can be achieved by the packet when it arrives at its destination.

Therefore, it is relatively important for each of the protocols discussed so far to have a common understanding of each other's commands and parameters. Only then will QoS agreements in a network be communicated to neighbouring networks, thus assuring that packets in a particular flow will receive the same if not similar treatment, in terms or delays, delivery rates or latency, across the different networks.

CONCLUSION

As multimedia communications become increasingly prevalent, it becomes important to deal with special conditions of multimedia traffic and provide for these special conditions with the architecture of an IP router/switch and its management and control. There is an opportunity to reconcile the perspectives of the computing and communication communities using integrated network architectures that support quality of service requests, fulfilment, control and continuity between different networks, and the joint allocation of computer and communication resources. The thrust of the IP community is towards supporting very few differentiated services, with different treatment for aggregation of traffic in each service class. One such research direction is to use MPLS in a trunk dedicated to a particular service class and RSVP for setting up and maintaining these trunks, the ultimate goal being the provision of integrated end-to-end QoS in today's Internet.

REFERENCES

Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., & McManus, J. (1999). *Requirements for traffic engineering over MPLS.* IETF RFC 2702.

Baker, F., Krawczyk, J., & Sastry, A. (1997). *RSVP management information base.* IETF RFC 2206.

Biswas, J., Lazar, A.A., Huard, J.F., Lim, K., Mahjoub, S., Pau, L.-F., Suzuki, M., Torstensson, S., Wang W., & Weinstein, S. (1998). The IEEE P1520 dtandards initiative for programmable network interfaces. *IEEE Communications Magazine,* (10), 64-70.

Black, U. (2002). *MPLS and label switching networks* (2nd ed). Saddle River, NJ: Prentice Hall.

Braden, R., Clark, D., & Shenker, S. (1994). *Integrated services in the Internet architecture: an overview.* IETF RFC 1633.

Callon, R., Doolan, P., Feldman, N., Fredette, A., Swallow, G., & Viswanathan, A. (1997). *A framework for multiprotocol label switching*. IETF RFC 2702.

Fineberg, V., Chen, C., & Xiao, X. (2002). An end-to-end QoS architecture with the MPLS-based core. *2002 IEEE Workshop on IP Operations and Management*, 26-30.

Gozdecki, J., Jajszcyk, A., & Stankiewicz, R. (2003). Quality of service terminology in IP networks. *IEEE Communications,* (3), 153-59.

Grossman, D. (2002). *New terminology and clarifications for diffserv.* IETF RFC 3260.

Guerin, R., Gan, D.H., Kamat, S., Li, T., & Rosen, E. (1997). *Setting up reservations on explicit paths using RSVP*. IETF RFC 2430.

Rosen, E. C., Viswanathan A., & Callon A. (2001). Multiprotocol Label Switching Architecture. IETF RFC 2430.

Stephenson, A. (1998). Diffserv and MPLS: A quality choice. *Data Communications*, November, 73-77.

Xiao, X., & Ni, L.M. (1999). Internet QoS: A big picture. *IEEE Network,* (2), 8-18.

Yoram, B., Binder, J., Blake, S., Carlson, M., Carpenter, B.E., Srinivasan, K., Davies, E., Ohlman, B., Verma, D., Wang, Z., & Weiss, W. (1999, February). A framework for differentiated services. *Differentiated Services Working Group.* Internet draft, retrieved April 27, 2002 from *http://draft-ietf-diffserv-framework-02.txt*

Zhang, L., Deering, S., Estrin, D., Shenker, S., Zappala, D. (1993). RSVP: A new resource ReSerVation protocol. *IEEE Network, 7*(5), 8-18.

KEY WORDS

Behaviour Aggregates (BA): A BA is a set of packets going in one direction of a link that exhibit similar QoS characteristics.

Differentiated Services (DiffServ): Framework where network elements give preferential treatment to classifications identified as having more demanding requirements. DiffServ provides QoS based on user group needs rather than traffic flows.

Flow: A set of packets associated with a single application and that share common requirements.

Forwarding Equivalence Class (FEC): A group of network packets forwarded in the same manner (e.g., over the same path, with the same forwarding treatment). A forwarding equivalence class is therefore the set of packets that could safely be mapped to the same label. Note that there may be reasons that packets from a single forwarding equivalence class may be mapped to multiple labels (e.g., when stream merge is not used).

Integrated Services (IntServ): Architecture where network resources are apportioned according to an application's QoS request, subject to bandwidth management policy and focused on individual flows.

Label: A short, fixed-length, physically contiguous, locally significant identifier used to identify a stream.

Label Swapping: A forwarding paradigm allowing streamlined forwarding of data by using labels to identify streams of data to be forwarded.

Per-Hop-Behaviour (PHB): Externally observable forwarding behaviour applied to a behaviour aggregate at a node.

Quality of Service (QoS): A measurable level of service delivered to network users, which can be characterized by packet loss probability, available bandwidth and end-to-end delay.

Resource Reservation Protocol (RSVP): A tool for prevention of congestion through reservation of network resources.

Service Level Agreement (SLA): Service contract between a service provider and its customer that defines provider responsibilities in terms of network levels (throughput, loss rate, delays and jitter) and times of availability, method of measurement, consequences if service levels aren't met or the defined traffic levels are exceeded by the customer, and all costs involved.

Stream: An aggregate of one or more flows, treated as one aggregate for the purpose of forwarding in layer-2 or layer-3 nodes (e.g., may be described using a single label). In many cases, a stream may be the aggregate of a very large number of flows.

ENDNOTES

1 The specific mapping of the drop precedence with ATM and Frame Relay is defined in RFC 3270.

2 DS-LER means a DiffServ capable LER.

3 Two things may happen here: 1. The packet came from another DiffServ domain, so it already contains a DSCP in its IP header; in this case, the classifier will only need to use that DSCP to map to the MPLS label. 2. If the packet came from a non-DiffServ domain, then the classifier will need to look at various header information (maybe both layers 3 and 4) to decide which BA this packet belongs to and then assign the corresponding DSCP/label.

4 The signalling protocol for MPLS is either Label Distribution Protocol (LDP), or an extension of RSVP.

Reliability Issues of the Multicast-Based Mediacommunication

Gábor Hosszú
Budapest University of Technology and Economics, Turkey

INTRODUCTION

Multimedia applications generally support the one-to-many group communication way. Multicasting decreases communication costs for applications that send the same data to multiple receivers. Table 1 summarizes the types of communication among hosts.

Currently, there is a fast increasing need for scalable and efficient group communication. The multicast theoretically is optimal for such purposes. Therefore, multicast technology is an emerging media dissemination technology instead of the traditional unicast communication. It has two important types: the network-level, namely IP-multicast, and the application-layer host-multicast. In the former, data packets are delivered by the IP protocol from one host to many hosts that are members of a multicast group. The routers run an IP-multicast routing protocol to construct a multicast tree. Along this tree, the data is forwarded to each host. Special IP addresses (224.0.0.0-239.255.255.255 address range) that define multicast channels and do not belong to given hosts are used. In case of Application-Layer Multicast (ALM), the hosts use unicast IP delivery, and the routers do not play any special role.

Reliability is one of the most important features of all multimedia applications, independent from the used multicast technology. This requirement is especially critical in the case of multicast, where, because of the large volume of data to be transferred, the correction or resending of lost data will be even more difficult in time.

In multicast technology, the maintenance of group membership information is also an important question from the point of view of the robustness of the so-called multicast delivery tree. In the case of an IP-multicast, the root of the tree is the sender, the leaves are the receivers and the intermediate nodes are the routers. In the following, the reliability properties of different multicast technologies will be reviewed.

RELIABLE IP-MULTICAST

The IP-multicast itself cannot guarantee any reliability, according to the well-known best-effort delivery of the IP network. To increase the reliability for the data distribution or interactive media applications, reliable transport protocols are necessary. However, unicast TCP does not support multicast, and the UDP does not provide any reliability. For this reason, additional multicast transport protocols are used to achieve the required level of reliability (Hosszú, 2001). The protocol stack of the reliable IP-multicast is presented in Figure 1.

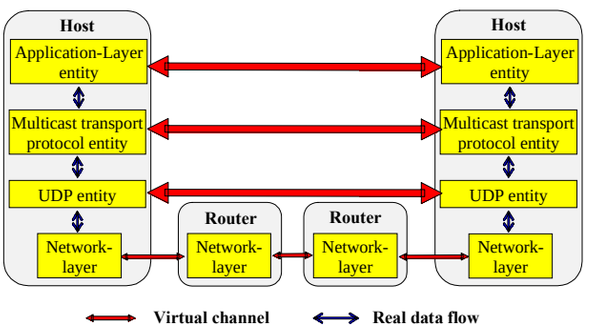
The various media applications, as the distributed collaborative multimedia systems, data dissemination tools and real-time media streaming software, require

Table 1. Possible types of communication among hosts

Type	Name	Description
point-to-point	**unicast**	One host communicates with another.
point-to-multipoint	**multicast**	One host (sender) sends data to a group of hosts; the sender sends data only once and every member of the group will receive the data.
multipoint-to-multipoint	**multipoint multicast**	In a communication session more than one sender exists, which independently sends data to every member of the group.
multipoint-to-point	**concast**	Every member of the group sends data to only one host.
point-to-everypoint	**broadcast**	One host sends data to every host.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Location of the multicast transport in the protocol stack



different multicast transport protocols for optimal performance. The multicast transport protocols have many different property attributes of data delivery, such as flow control, congestion control, data and time reliability, packet ordering, state control, acknowledgement control, scalability of the repair requests and so forth. These attributes can be represented by a selected set of protocol parameters. Each protocol parameter describes different reliability mechanisms for the same delivery attribute. Such a protocol parameter is, for instance, the repair method, which can get such values as the retransmission, forward-error correction, interleaving or different ways of the local receiver-based repairs. Another parameter is the acknowledgement type, the possible values of which may be tree-based, ring-based or a simple direct form.

To improve multicast reliability, the optimization of these protocol parameters is necessary. However, to apply any appropriate mathematical optimization method at the selection of the protocol parameters mentioned above, a linearly independent (or *orthogonal*) set of parameters must be applied. To do this, a hyperspace of the parameters is created where all transport protocol corresponds to one point of this space. The optimization procedure means finding the most suitable point on this space to provide the best performances of multicast. The modeling procedure based on the introduced protocol parameter set is presented on some examples. The strengthness of this orthogonality may be weakened, as discussed later.

The possible values of protocol parameters (which are the types of various mechanisms as the components of the transport protocols) are the realizations of the protocol functionalities. Table 2 shows a possible set of 31 different protocol parameters and their classification into categories. These parameters represent the well-known reliable mechanisms of transport protocols. The details of these mechanisms are described in the pertinent literature (e.g., Adamson et al., 2004).

For an individual application, protocol parameters get actual values. To optimize a transport protocol, the optimal point should be found in the 31-dimensional *hyperspace of the protocol parameters*. The optimization procedure can be executed easily if the applied protocol parameters are orthogonal to each other. Orthogonality means that any of them can be changed independently from the others. Since the selection of the applied protocol parameters is very important, the task is to obtain a complete set of protocol parameters that can be taken as orthogonal. For the current set of 31 protocol parameters, orthogonality is not completely satisfied, but because the importance of different protocol parameters are

Table 2. The 31 protocol parameters

Category	Protocol parameters
Data traffic control	Transmission way, transmission direction, congestion prevention, flow control
Delivery control	Data accuracy, time limitation, scheduling, updating, ordering
Feedback management	Acknowledgement types, feedback addressee, election of the designated host, state control, feedback control, way of providing feedback
Repair management	Request way, repair method, repair source, repair selection, way of sending repair, repair scoping, repair control
Session management	Session control, floor control, session membership control, locus of control, scalability, group stability
Network demand	Bandwidth demand, network heterogeneity, direction dependency

highly different, it can be used. The parameters that depend slightly on each other are called quasi-orthogonal parameters. Its practical definition is that changing one of their values does not involve change in the value of any other; however, some combinations of parameter values can be inefficient. An example of negligible dependency is the relation between the parameters feedback control and feedback addressee, where the influence of modifying the actual value of feedback control from structure-based to timer-based – and the optimal value of feedback addressee from intermediate host to every member – may practically be ignored. A quasi-orthogonal subset of the protocol parameters and their possible values are presented in Table 3.

The protocol parameter flow control means the prevention of the receivers against the overload. The data accuracy is reliable – if there is mechanism in the protocol for loss recovery and it is atomic – if the protocol provides exactly the same data for all receivers. Feedback addressee is the host, which receives the positive or negative acknowledgements. The state control parameter defines the responsibility for loss detection. Feedback control means the prevention of the feedback addressee host against the feedback implosion. The way of sending repair determines the applied transporting mechanism. The scope of repair describes the responsibility of the repairing source, which can be global or local. If its responsibility is global, it can send packets to every member, to a subset of the whole group (secondary group) or even to individual members only. Finally, the parameter session membership control describes the handling method of hosts that want to join the session. This protocol parameter is explicit if the members of the session are registered, and implicit if they are not.

The selected parameters are quasi-orthogonal only, since there are trivial cases where the orthogonality cannot be satisfied. For example, if there is no feedback, then protocol parameters feedback addressee and feedback control will be none. However, if the feedback control parameter is none, the feedback addressee is not obviously none, as Table 4 shows, where the actual protocol parameter values of four transport protocols are displayed. The presented transport protocols are Transport protocol for Reliable Multicast (TRM) (Sabata et al., 1996), Log-Based Reliable Multicast (LBRM) protocol (Holbrook et al., 1995), Light-weight Reliable Multicast Protocol (LRMP) (Liao, 1998) and Reliable Adaptive Multicast Protocol (RAMP) (Koifman & Zabele, 1996). All of the presented multicast transport protocols have receiver-based state control. Naturally, in order to describe a certain protocol mechanism more specifically, the usage of additional, secondary parameters may be necessary which make the fine-grained tuning of each protocol mechanism.

The reason of the orthogonality of the protocol parameters is that the background mechanisms are independent from each other.

Other multicast transport protocols can be defined by the previously mentioned and additional protocol parameters, including: Real-Time Transport Protocol (RTP) for real-time applications (Schulzrinne et al., 2003) and NACK-Oriented Reliable Multicast Protocol (NORM) for bulk data transfer (Adamson et al., 2004).

To carry out a correct optimization procedure on the appropriately selected protocol parameters, a confidential simulation should be applied in order to get statistically realistic results for multicast data

Table 3. A selected subset of protocol parameters

Protocol parameter	Values
Flow control	Window-based, rate-based, multigroup multicast, receiver give-up, none
Data accuracy	Reliable, atomic, non-reliable
Feedback addressee	Original source, intermediate host, every member, none
State control	Sender-based, receiver-based, shared, none
Feedback control	Structure-based, timer-based, representatives-based, rate-based, none
Way of sending repair	Unicast, multicast, none
Scope of repair	Global, global to secondary group, global to individuals, local, none
Session membership control	Explicit, implicit, none

Table 4. Actual values of protocol parameters in four transport protocols

Protocol parameter	TRM	LBRM	LRMP	RAMP
Flow control	Window-based	None	Rate-based	Rate-based
Data accuracy	Reliable	Reliable	Reliable	Non-reliable
Feedback addressee	Original source	Intermediate host	Every member	Original source
State control	Receiver-based	Receiver-based	Receiver-based	Receiver-based
Feedback control	Timer-based	Representatives-based	Timer-based	None
Way of sending repair	Multicast	Unicast or Multicast	Multicast	Unicast or Multicast
Scope of repair	Global	Global to individual member or local	Local	Global to individual members or local
Session membership control	Implicit	Implicit	Implicit	Explicit

transfer. Using an appropriate simulator, an optimized transport protocol can be synthesized, satisfying the requirements of a certain media application. This means that by a mathematical method an optimal point in the hyperspace of the protocol parameters can be found.

PEER-TO-PEER NETWORKS

The host can create a virtual network overlaid on the Internet, in which the members are peering entities – they are responsible for maintaining the overlay network. Such a network is called Peer-to-Peer (P2P), which differs from the widely used client-server model. P2P networks are created by host-multicast (also called end-host multicast or ALM) technology, where the nodes are responsible for the multicast delivery, not the routers. The main difference between ALM and IP-multicast is that at ALM the hosts use unicast for the data transmission between themselves, and the multiplication points of the multicast tree are the nodes, not the routers. In order to create the multicast tree, the hosts run a so-called host-multicast routing protocol. Unlike IP-multicast, ALM requires no infrastructure support and can be easily deployed in the Internet.

The basic idea of ALM is that the multicasting functionality is implemented at the application layer, at the end-hosts instead of the routers. In the IP-multicast, data packets are replicated at routers inside the network; however, in an ALM data packets are replicated at end-hosts. Virtually, the end-hosts form an overlay network, and the goal of ALM is to construct and maintain an efficient overlay for data delivery.

The special type of the ALM is the hybrid approach, which goals to reach a ubiquitous multicast. One design requirement of it is that it should be deployable on the current Internet without any change in the operating systems, routers or servers. The other design requirement is the compatibility with IP-multicast, which means that it should use IP-multicast where available.

A special field of the ALM is multicasting in a wireless environment. Such an environment for multicast is the mobile ad-hoc network (MANET) (Corson & Macker, 1999). MANET consists of a dynamic collection of nodes with possible rapidly changing multi-hop topologies that are composed of low-bandwidth wireless links. The majority of nodes use batteries; therefore, routing protocols have to limit the amount of control information forwarded between nodes. The applications for MANET technology are in areas where fast deployment and dynamic reconfiguration are necessary and a wired network is not available. Multicast technology improves the efficiency of the wireless link when sending multiple copies of packages.

Group membership maintenance is a serious problem, even in traditional wired networks; but currently, the possible nodes are increasingly mobile. Therefore, multicasting in a MANET means a new challenge for the multicast technology (Kunz & Cheng, 2002).

ALM protocols organize group members into two topologies: the control topology and the data topology (Banerjee et al., 2002). Members periodically ex-

change refresh messages to maintain the control topology. The data topology is usually a subset of the control graph and defines the data path for a multicast packet to be transmitted. The data topology is a distribution tree, while the control graph (sometimes called mesh) has richer connectivity between member-hosts.

Depending on the sequence of construction of the control and data graphs, the various ALM approaches belong to the following classes: mesh-first, tree-first and implicit. Their main properties are shown in Table 5.

From the viewpoint of the control overhead, the topological distribution of the members is important. In a dense mode scenario, where the huge part of the total sum of the hosts are members in a certain network domain, the propagation delays of the refresh messages are small and the maintenance of the overlay is not costly. In case of the so-called sparse mode, where the members are sparsely distributed throughout the Internet, the control message overhead could be much higher. In this scenario, the inherent instabilities of the Internet can play a serious role in the degradation of the performance of the ALM protocol.

A typical ALM solution is the Your Own Internet Distribution (YOID), which supports a variety of applications, ranging from file transfer to real-time multiparty media tools, network news, streaming broadcast and bulk mail distributions (Francis, 2000).

A new node joins a YOID group through a rendezvous node. Once a node has selected an appropriate parent from those offered by the rendezvous node, the new node joins the shared tree that is then used to transfer data.

STF: A NOVEL ALM ROUTING PROTOCOL

As an example of the current searching works in the field of ALM, a new concept of modeling relative density of members is called bunched mode. It constitutes a typical multicast scenario, where there are many interested hosts in certain institutes and these institutes are relatively far from each other. The members of a multicast group are locally in dense mode; however, these spots are far from each other, globally, their situation is similar to the sparse mode. The bunch can be a Local-Area Network (LAN) or just an Autonomous System (AS). This situation is typical when the collaborative media application has a special topic. That is why this model of communication is called Thematic Multicast Concept (TMC).

The group-member hosts in a bunch locally elect Designated Members (DMs). The DMs calculate the shortest unicast IP tunnels among them, and they exchange their IP addresses and shortest unicast paths. In such a way, all of them know all the possible shortest unicast paths and calculate the same topology of the inter-bunch IP tunnels. This mechanism, called Shortest Tunnel First (STF), is similar to the Multicast Open Shortest Path First (MOSPF) routing in network level (Moy, 1994).

The STF does not require any global rendezvous point for creating the inter-bunch delivery tree. However, suppose that there is only one source per group and constructs unidirectional tree. A typical TMC scenario is shown in Figure 2.

The list of the DMs is maintained by the source application, and the new DMs register here and get the

Table 5. Classes of ALM methods

ALM classes	Properties	Example protocols
Mesh-first	Group members organize themselves into the overlay mesh topology, then run the Application-Layer routing protocol to construct a source-rooted data tree. It is efficient for small-sized groups.	Narada (Chu et al., 2001) STF (see article section)
Tree-first	The Application-Layer routing protocol constructs a multicast tree directly, then each member discovers some other, non-neighboring members and creates control links to these hosts. It is well suited for data transferring applications that need high bandwidth, but not efficient for real-time purposes.	HMTP (Zhang et al., 2002) Yoid (Francis, 1999)
Implicit	The mesh and the tree are simultaneously defined by the protocol. It scales well to multicast groups with large number of members.	CAN (Ratnasamy et al., 2001) NICE (Banerjee et al., 2002) Scribe (Castro et al., 2002)

copy of the current list. Periodically all DMs send IP packets to every other DM and send a keep-alive message to the source if a DM is not available. If a DM is not reported to the source, it is deleted from the list. The source periodically sends the list to the DMs. If the source is not available, the group state is timed out in the DMs.

The STF protocol constructs an almost similar optimal tree than the IP-multicast. However, it does not require any inter-domain multicast routing mechanism in the routers. It belongs to the mesh-first class. It is optimal for relatively small groups, but due to the TMC method, the topological size of the group does not limit its scalability.

CONCLUSION

The two different types of multicast technology, IP-multicast and ALM, were described. The reliability of the IP-multicast communication can be increased by the multicast transport protocols that use various protocol mechanisms. The systematic modeling of the multicast transport protocol mechanisms was presented, which were selected to be linearly independent from each other. This novel protocol parameter set was shown and the modeling procedure was demonstrated. A multi-dimensional hyperspace was stated as a mathematical model, where every transport protocol is represented with an individual point.

The new routing protocol of the ALM is the STF, which also was introduced. It can be used in a typical scenario that was modeled with the TMC communication model.

REFERENCES

Adamson, B. et al. (2004). NACK-Oriented Reliable Multicast protocol (NORM). *IETF, draft-ietf-rmt-pi-norm-09.txt*, work in progress. Available at *http://www.ietf.org/internet-drafts/draft-ietf-rmt-pi-norm-09.txt*

Banerjee, S. et al. (2002). Scalable Application-Layer Multicast. *ACM SIGCOMM*.

Castro, M. et al. (2002). Scribe: A large-scale and decentralized Application-Layer Multicast infrastructure. *IEEE Journal on Selected Areas in Communications (JSAC), 20*(8).

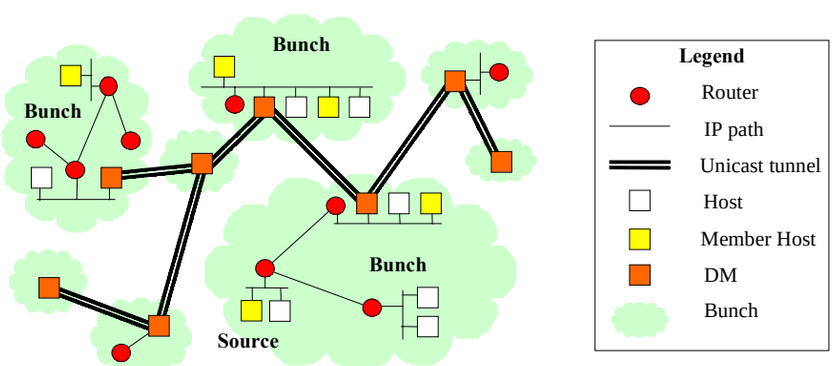
Chu, Y.-H. et al. (2001). Enabling conferencing applications on the Internet using an overlay multicast architecture, *ACM SIGCOMM*.

Corson, S., & Macker, J. (1999). Mobile ad hoc networking (MANET): Routing protocol performance issues and evaluation considerations. *IETF RFC 2501*.

Francis, P. (1999). Yoid: Extending the multicast Internet architecture. White paper, retrieved from *www.aciri.org/yoid*

Holbrook, H.W. et al. (1995). Log-based receiver-reliable multicast for distributed interactive simulation. *ACM SIGCOMM*, 328-341.

Figure 2. A typical TMC scenario



Hosszú, G. (2001). Introduction to multicast technology. In S.M. Rahman (Ed.), *Multimedia networking: Technology, management & applications* (pp. 369-411). Hershey: Idea Group Publishing.

Koifman, A., & Zabele, S. (1996). Ramp: A reliable adaptive multicast protocol. *IEEE INFOCOM,* San Francisco, CA, 1442-1451.

Kunz, T., & Cheng, E. (2002). On-demand multicasting in ad-hoc networks: Comparing AODV and ODMRP. *ICDCS'02*.

Liao, T. (1998). Light-weight reliable multicast protocol. Technical Report, INRIA, Le Chesnay Cedex, France.

Moy, J. (1994). Multicast extensions to OSPF. *IETF Network Working Group RFC 1584.*

Ratnasamy, S. et al. (2001). A scalable content-addressable network. *ACM SIGCOMM.*

Sabata, B. et al. (1996). Transport protocol for reliable multicast: TRM. *IASTED International Conference on Networks*, 143-145.

Schulzrinne, H. et al. (2003). RTP: A transport protocol for real-time applications. *IETF Network Working Group RFC 3550*.

Zhang, B. et al. (2002). Host multicast: A framework for delivering multicast to end users. *IEEE INFOCOM*.

KEY TERMS

Ad-Hoc Network: This is a special type of computer network, where \communication does not require any fixed computer network infrastructure (e.g., it does not need a router); the nodes communicate directly with each other without access points. In *host-multicast* (*see below*), mobile peering hosts a construct ad-hoc network.

Client-Server Model: A communicating way, where one host has more functionality than the other. It differs from the ***P2P network*** (*see below*).

Host-Multicast: Application-layer multicast technology that does not require any additional protocol in the network routers, since it uses the traditional unicast IP transmission. Its other names are ***end-host multicast***, ***application-layer multicast***.

Host-Multicast Routing Protocol: The members of the hosts construct a delivery tree using similar algorithms than the IP-multicast routing protocols.

IP-Multicast: Network-level multicast technology that uses the special Class-D IP-address range. It requires multicast routing protocols in the network routers.

IP-Multicast Routing Protocol: To forward the multicast packets, the routers have to create multicast routing tables using multicast routing protocols.

Multicast Transport Protocol: To improve the reliability of the multicast delivery, special transport protocols are used in addition to the unreliable **U**ser **D**atagram **P**rotocol (**UDP**).

Peer-to-Peer (P2P) Network: A communication way where each node has the same authority and communication capability. They create a virtual network, overlaid on the Internet. Its members organize themselves into a topology for data transmission.

Reliability: The improved quality of data transmission. Different types of reliability exist, including data accuracy or real-time delivery.

Re-Purposeable Learning Objects Based on Teaching and Learning Styles

Abtar Kaur
Open University of Malaysia, Malaysia

Jeremy Dunning
Indiana University, USA

Sunand Bhattacharya
ITT Educational Services, Inc., USA

Ansary Ahmed
Open University of Malaysia, Malaysia

INTRODUCTION

Web-based distance learning is hampered in many cases by a failure to deliver material in a manner consistent with the ways in which students learn and instructors teach best in traditional environments (Samorski, 2002). Excellent teachers are successful because of the ways in which they mediate content and place the content within the context of the subject matter. It is not the specific content or images the successful teacher presents, but rather the manner in which they are presented and framed within the scope of the topic area. Excellent teachers teach by presenting the content and then providing the students with substantive opportunities to apply the content to real-world problems in an effort to promote critical thinking on the part of the student. This is a highly interactive process with much information being transmitted between the student and the instructor. The interchange between the instructor and the student helps the student build a knowledge base with the assistance of the instructor's experience and expertise in the topic area. The exact nature of the interchange is not predetermined and depends to a great extent on the creativity and breadth of experience of the instructor. The successful instructor adjusts his or her interaction with the students to the learning styles best suited to them. How do we provide the learner with this important component of traditional classroom education in asynchronous distance education or technology-mediated traditional classes? Web-based instruction is rapidly becoming the preferred mode of distance education, and we must adapt our instructional interaction styles to this medium. Our students now expect more interactive and immersive materials in Web-based learning than that typically provided in the traditional classroom or correspondence distance education (Samoriski, 2002).

The TALON learning object system is a series of re-purposeable learning object templates based on styles of teaching and learning as described by Dunning et al. (2002). These Flash-based templates allow instructors to design and execute interactive learning objects in approximately 10% of the time required to create them from first principles, because the use of them requires little or no alteration of existing source code or writing of additional code (Abtar et al., 2004, Dunning et al., 2004). The fact that the learning objects are based on the successful learning styles experienced in the traditional classroom ensures that the student is both engaged and allowed to build a knowledge base about the content being covered.

BACKGROUND

The overall online course design process can be classified broadly into three phases: development, delivery, and results. The development phase is collaborative in nature where the actual course gets designed and constructed; the delivery phase is where the instructor interacts with the students via the online

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

course; and the third phase is where outcomes translate into learning competencies.

Retention and attrition issues in an online course are often attributed to the level of interest the course generates. The immersive nature of a course depends on its engaging features. Often, complex concepts or phenomena can be taught better through interactive models that encourage the student to explore and learn. Appropriate design of a distance education course delivered through suitable media and using befitting strategies enhances learning (Fennema, 2003). Designers of effective distance courses delivered through the Internet must consider the interactivity of the medium and employ it to enhance the instruction of the distance learner (Hirumi & Bemudez, 1996; Starr, 1997).

Learning Objects

Learning objects have been defined in a number of ways by many researchers. Some define learning objects as any visual feature that engages the student's attention (Wiley, 2000). Others require a certain degree of interactivity for material to be considered a learning object (Wisconsin Online Resource Center, 2003). For the purposes of this discussion, it will be assumed that learning objects must be interactive to be considered true learning objects. The National Learning Infrastructure Initiative defines learning objects as "modular digital resources, uniquely identified and metatagged, that can be used to support learning." The common threads in all of these definitions are summarized in Table 1.

Learning Styles

Although most educational researchers agree that individual differences in the ways in which students learn play a role in learning, there is little agreement on the nature of the different ways students learn. There is little agreement even on the terminology applied to ways in which students learn. Terms such as learning styles, cognitive styles, learning preference, learning strategies, and learning modalities are used to describe the same basic phenomenon—the manner in which students learn. Researchers use these terms almost interchangeably; however, learning style is the most commonly used term and will be used here. Learning style is generally accepted to be a student's existing learning strengths or preferred manner of learning (Kaplan & Kies, 1995).

Marineti (2003) and De Bello (1990), among others, have classified learning style as a subset of cognitive style. Others (Morse, 2003) feel that learning style encompasses cognitive style. The majority of researchers agree that individuals have different learning styles and that an individual modality of learning is not equally effective for all learners (Sims & Sims, 1995). Sadler-Smith (1997) identified four categories of learning styles: cognitive personality elements, information processing style, instructional preferences, and approaches to study.

A number of assessment tools and quantitative indices have been developed to define an individual's learning and cognitive styles. The early seminal work includes the Myers-Briggs Type Indicator, the Cognitive Preference Test (Messick, 1984), the Cognitive

Table 1. Attributes of learning objects

Learning objects help students:	**Identify features and processes interactively through visual learning.**
Learning objects allow students:	**To solve real-world problems by immersion in an interactive scenario, based on the content they are covering.**
Learning objects provide students:	**With the opportunity to make and interpret empirical observations in a digital environment that simulates a real-world situation.**
Learning objects help students:	**Develop critical thinking skills and, in some cases, verbal skills.**
Learning objects help students:	**Realize that they have achieved certain learning benchmarks and build confidence in their mastery of the content.**

Style Profile (Kuckinskas, 1979), and the Learning Style Inventory (Kolb, 1976). More recent reviews of learning and cognitive styles includes Dunn (2003) and Fennema (2003).

Collaborative learning is an important learning style that has so far been restricted for the most part to the traditional classroom, where it has been a successful learning strategy. Recent work by Hislop, Hassell, and Wiedenbeck (2003) and Hildebrandt (2003) has demonstrated that collaborative learning can be effectively executed in an online environment.

RE-PURPOSEABLE LEARNING OBJECT TEMPLATES

One of the struggles faced in distance education and technology-mediated instruction is providing interactive and highly experiential learning exercises. Learning objects are useful in this setting because they allow the student to use the content learned in a particular part of a course and (1) demonstrate mastery of the content; (2) apply that knowledge to solving a problem; and (3) use the content in a critical thinking exercise that allows the student to place the content within the context of the larger course topic.

Learning Objects may be problematic in several ways. First, they require some multimedia programming, and, therefore, they are beyond the abilities of typical instructors, who may be capable only of creating course support materials within simple authoring tools such as PowerPoint. Additionally, these objects are usually created from first principles each time, and the cost of providing substantive interactive learning objects may be prohibitive. In part, the cost is related to the fact that the programmers and instructional designers may not know a great deal about the subject matter, and the instructors may know little about multimedia design and programming. Both groups, therefore, operate within their own areas of comfort, and there is little real communication outside of those areas.

It is clearly impractical to teach each instructor about multimedia programming and teach each programmer about specific subject matter areas. What is practical is to define learning objects in terms of the styles in which we teach and learn. All of us understand how we learn, and there are a finite number of learning styles. Educators tend to teach to those learning styles, consciously or not, because they know from experience that teaching styles that are linked to the ways in which students learn are most effective. If we were to define learning objects more in terms of the teaching and learning styles that the objects utilize, and less in terms of the specific content or programming strategy, programmers and instructors could more clearly understand each other and the role each plays in the design process. The instructor can be more involved in the design of the learning objects, if the objects are defined in terms of a context (teaching and learning styles) that he or she can understand. Developing a common language of design cuts the cost of developing individual learning objects, because it reduces the number of modification cycles between the subject matter expert and the programmer. The cost remains high, however, if each object is designed from scratch. It also allows designers and programmers to move from content area to content area using the same nomenclature and design principles, because teaching and learning styles are independent of topic area.

If learning objects are defined in terms of a limited number of teaching and learning styles, they are independent of content area to a great extent. Therefore, we should be able to create templates for learning objects that are based on learning or teaching styles. The templates would be designed so that they could be reprogrammed for any content area at minimal expense. This would allow instructors to design learning objects for their courses using most of the existing code for the template. A multimedia programmer would then insert the graphic and text elements required to complete the learning object in the design executed by the instructor. In most cases less than 5% of the code for the template would need to be rewritten each time the learning object is reconfigured.

The TALON learning object suite, developed by Arjuna Multimedia and further developed in conjunction with Open University of Malaysia, is a set of 39 re-purposeable learning object templates based on styles of teaching and learning that are designed to allow instructors and designers to create substantive learning objects without changing any of the source code. The instructor or designer can use the templates to design a new learning object without writing or changing any source code. The templates are simple enough that instructors with little or no

programming experience can create their own learning objects. For more information about the TALON Learning Object Suite, visit http://www.arjunamultimedia.com/Website/encycl.htm.

The TALON system does not rigorously follow any single model of learning styles described previously, and is based on the learning styles and strategies as defined by a group of more than 30 university and high school instructors. These strategies are based more on the instructors' experiences than on any particular theoretical model of teaching styles and strategies. Because of this, they combine the features of many of the models described previously.

REFERENCES

Abtar, K., Dunning, J., Harvinder, K, & Halimatolhanin, M. (2004). How reusable are learning object templates: A case study. *Proceedings of the 4th Pan Commonwealth Forum*, Dunedin, New Zealand.

De Bello, T.C. (1990). Comparison of eleven major learning style models: Variables, appropriate populations, validity of instrumentation, and the research behind them. *Reading Writing and Learning Disabilities, 6*, 203-222.

Dunn, P. (2003, December 2-6). Adapting e-learning for global environments: What have we learned so far? *Online Educa Berlin.*

Dunning, J., Cunningham, D., Kaur, A., & Vidalli, A. (2003). Artificial intelligence and learning objects optimizing the pedagogical impact of online learning. *Proceedings of the 9th Sloan-C Conference on Asynchronous Teaching and Learning*, Orlando, Florida.

Dunning, J., et al. (2003). Re-purposeable learning objects linked to teaching and learning styles. *Proceedings of the EISTA 03 International Conference on Education and Information Systems*, .

Dunning, J., et al. (2004). Technology is too important to leave to technologists. *Journal of Asynchronous Learning Networks, 8*(3), 14, 23-28.

Fennema (2003). Preparing faculty members to teach in the e-learning environment. In S. Reisman (Ed.), *Electronic learning communities* (pp. 239-269). Greenwich: Information Age Publishing.

Hilldebrandt, M. (2003). Cooperative e-learning and trancultural communication. *Online Educa Berlin*. Berlin.

Hirumi, A., & Bemudez, A. (1996). Interactivity, distance education, and instructional systems design converge on the information superhighway. *Journal on Computing in Education, 29*(1), 1-16.

Hislop, G., Hassell, L., & Wiedenbeck, S. (2003). Participant activity in online classes: Patterns and implications. *Proceedings of the 9th Sloan ALN Conference*, Orlando, Florida.

Kaplan, E., & Kies, D. (1995). Teaching styles and learning styles: Which came first. *Journal of Instructional Psychology, 22*, 29-33.

Kolb, D. (1976). *Learning style inventory: Technical manual*. Boston: McBer Press.

Kuckinskas, G. (1979). Whose cognitive style makes the difference? *Educational Leadership, 18*, 269-271.

Marinetti, A. (2003). The promise of learning style reuseability: Myth or reality? *Online Educa Berlin*. Berlin.

Messick, S. (1984), The nature of cognitive styles: Problems and promise in educational practice. *Educational Psychologist, 19*, 59-74.

Morse, K. (2003). The multicultural e-classroom: Learning, satisfaction, and faculty issues. *Proceedings of the 9th Sloan-C Conference on Asynchronous Teaching and Learning,* Orlanda, FL.

Sadler-Smith, E. (1997). Learning style: Framework and instruments. *Educational Psychology, 17*, 51-63.

Samoriski, J. (2002). *The Internet, children, and education: Issues in cyberspace, communication, technology, law, and society on the Internet frontier.* Boston: Allyn and Bacon.

Sims, R., & Sims, S. (1995). *The importance of learning style*. Westport, CT: Greenwood Press.

Starr, R. (1997). Asynchronous learning networks as a virtual classroom. *Communications of the Association for Computing Machinery, 40*(9), 44-49.

Wiley, D. A. (2000). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In *The instructional use of learning objects: Online version.* New York: D.A. Wiley.

Wisconsin Online Resource Center (2003). *http://www.wisc-online.com/Info/FIPSE% 20%20What%20is%20a%20Learning%20Object.htm*

Witmer, D. (1998). Introduction to computer-mediated communication: A master syllabus for teaching communication technology. *Communication Education, 47*.

KEY TERMS

Distance Learning: Learning in which the instructor and students are separated by time, distance, or both.

Flash: A multimedia authoring tool in which interactive learning objects may be created.

Interactivity: Occurs when a student works within a multimedia exercise in which the student and the program interchange information in order to complete the exercise.

Learning Objects: Interactive computer-based exercises in which a student utilizes critical thinking skills, achieves learning benchmarks, and displays mastery of content.

Learning Style: Generally accepted to be a student's existing learning strengths or preferred manner of learning.

Re-Purposeable Learning Objects: Learning objects that are designed as templates that can be reconstructed to serve new learning objectives.

Technology-Mediated Instruction: Learning that is aided or entirely accomplished through the use of computer-based technology.

A Risk–Control Framework for E–Marketplace Participation

Pauline Ratnasingam
Central Missouri State University, USA

E-MARKETPLACES – INTRODUCTION

The increasing trend in the use of Internet-based e-marketplace applications has created tremendous opportunities for businesses to manage effective supply chain management. Forrester Research predicts that the growth and use of e-marketplaces may reach US $7 to $10 trillion by the year 2005 and may account for 53% of all online businesses by the year 2005. White and Daniel (2003) describe e-marketplaces as Web-based systems that enable automated transactions, trading, or collaboration between business partners. According to Bakos (1998), an electronic marketplace is an interorganizational system that allows participating buyers and sellers to exchange information about processes, products, and services. This study aims to examine risks in e-marketplaces. We identify four types of risks: economic, technological, implementation, and relational risks in seven e-marketplace firms from a cross-section of different industries. We then present the control measures in the responses that the seven firms enforced in order to reduce and manage their risks. The contribution of this study is the development of a risk-control framework based on the findings for e-marketplace participation.

BACKGROUND INFORMATION OF THE E-MARKETPLACE FIRMS

Previous research has identified different types of e-marketplaces, including buyer-driven, seller-driven, vertical, horizontal, and enabling technologies that provide online buying services, auctions, functional exchanges, and net markets (Christiaanse & Markus, 2002, 2003; Kaplan & Sawhney, 2000; Lenz et al., 2002). Bailey and Bakos (1997) suggest these key roles of e-marketplaces: matching buyers and sellers, aggregating and facilitating buyers' demands, sellers' product, and acting as an agent of trust. Similarly, Kaplan and Sawhney (2000) classified B2B marketplaces as a two-by-two scheme considering dimensions on what firms purchase (manufacturing inputs or operation inputs) as well as how they purchase (spot buying or systematic buying).

Seven e-marketplace firms participated in this study from a cross-section of industries; namely, automotive, aerospace and defense industry, chemicals, construction, energy, agriculture, and plastics. First, Covisint, representing a buyer-driven e-marketplace in the automotive industry, has grown dramatically, and in January 2003, they had 76,000 registered trading partners. Second, Exostar in the aerospace and defense industry joined forces with Boeing, Lockheed Martin, Raytheon, and BAE Systems, four of the world's leading aerospace and defense manufacturers, to streamline the highly sophisticated processes that aimed to achieve zero tolerance for failures. Third, ChemConnect, originally conceived as a bulletin board site in 1995, primarily deals with feedstock, chemicals, plastics, and other products. It advertises more than 9,000 members from 150 different countries. Fourth, Construction.com, owned by McGraw Hill, acts as a portal for construction operations like its Dodge analytical service, helps contractors estimate costs, and make bids. Fifth, Pantellos group is an open marketplace focused on utility and energy services. Many large utilities in North America, including Houston-based Reliant Energy Inc., formed Pantellos group in the year 2000 to create supply chain services. Sixth, Farms.com, an e-marketplace in the agriculture industry established by merging with Agribiz.net's eHarvest, grew to offer an online trading platform for a wide range of agricultural products including swine, beef, dairy, cattle, poultry, real estate, and crop protection. Finally, Omnexus, a plastics industry owned by a Dutch company, offers browser-based transaction software for customers to buy directly from suppliers.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

E-MARKETPLACE RISKS

Despite the growth and hype of e-marketplaces, uncertainties, vulnerabilities, and risks existed in e-marketplaces (Choudhury et al., 1998; Sims & Standing, 2002). Exposures to risks increased when disparate services were provided to trading partners. The new services were threatened by internal factors such as lack of standards, lack of regulations, and lack of secure systems. Furthermore, external factors such as the volatile online political sanctions, natural hazards, legal issues, environmental issues, and other political instabilities threatened the firms. Reshaur and Turner (2000) suggest that risks can be viewed as hazards, uncertain outcomes, or missed opportunities. We categorized risks in e-marketplaces as economic, technological, implementation, and relational, discussed as follows.

Economic risks are derived from increased transaction costs. Most of the independent e-marketplaces such as Aluminum.com, Ventro Corporations, Chemdex, and Promedix were unable to maintain their liquidity due to large manufacturers generating large volumes of transactions and negotiating with suppliers and vendors on their own, thereby saving transaction costs for themselves while ignoring the smaller suppliers. Similarly, Vertical Net Inc and SciQuest Inc have transformed their businesses to become software vendors, thereby avoiding charging fees for online transactions (Hicks, 2001; Segev et al., 1999).

Transaction costs consist of coordination cost (made up of search cost for finding the right supplier or buyer) and the cost for exchanging information. Furthermore, contracting costs include the cost of negotiation as well as legal and administrative costs incurred in creating an enforceable contract that satisfies both trading parties (Gulledge & Mason, 2000; Le, 2002; Premkumar, 2003). Sklar (2001) and suggest that as e-marketplaces continue to evolve, the key component for their survival will be their ability to sustain global e-commerce liquidity and efficiency through trust-based transaction and settlement solutions.

Technological Risks

Technological risks are derived from integration issues (i.e., incompatible applications) and security issues (i.e., the volatile Internet environment). Technological risks impact suppliers, as they are required to adopt different technological solutions provided by the buyers (also known as technology squeeze). New and untested applications created scalability, security, and availability issues (Vaidyanathan & Devaraj, 2003). Poor business practices create administrative threats such as password sniffing, data modification, spoofing, and repudiation. Furthermore, a variety of standards and operating procedures caused a lot of frustration and resistance among suppliers. Suppliers who have attempted to outsource their business applications, in the hope of reducing costs and maintaining profit margins, faced the risks of outsourcing their critical business processes to different firms with different procedures (Gulledge & Mason, 2000). Likewise, changes in the online fulfillment processes have posed technological risks, as products and services are needed almost in real time. Integration of real-time sales orders with the existing supply chain management and order fulfillment has caused trading partners to be exposed to pressure.

Implementation Risks

Implementation risks are derived from the lack of bargaining power due to relationship-specific investment. Suppliers have attempted to implement the same technological path as buyers, but it leads to the risk of continued investment in new technologies and additional integration costs. Operation risks are increased due to the lack of technical knowledge of the system and training (Premkumar, 2003; Wise & Morrison, 2000).

Due to a proliferation of different types of technological solutions, maintaining standards in the management and business processes has become challenging (Cuny & Richardson, 2001). The lack of uniform standards increases the exposure to risks when disparate services are offered. Risks associated with e-business derived from poor business practices arose from applying weak procedures in the software development process; having deficiencies in the e-business protocols leads to technology-related problems. Furthermore, implementation risks are derived when trading partners introduce information links that closely monitor changes in their customer base (Vaidyanathan & Devaraj, 2003).

Relational Risks

Relational risks are derived from the failure to address power-related issues among trading partners. Focus on dyadic relationships means that interdependencies among a series of related trading relationships (i.e., power) may be missed. E-marketplaces reduce the power of sellers because of transparency of prices. Imbalance of power leads to a lack of trust among trading partners concerned with falsifying e-documents. Similarly, buyers could form coalitions and counter the bias of supplier-owned systems by colleting the data and reformatting them to meet their own needs. Santos and Perogianni (2001) suggest that competition, standardization, and a lack of trust cause relational risks. Challenges in the competition include misuse of trading agreements, exchange of information, and restriction of freedom of suppliers and/or buyers to participate on equal terms.

FINDINGS – E-MARKETPLACE RISKS EXPERIENCED BY THE FIRMS

In this section, we discuss the findings from seven e-marketplace firms pertaining to the risks they experienced.

Economic Risks

Economic risks in the automotive industry arose when online parts for procurement led to reduced costs of information technology spending, as Covisint was applying new untested technologies that had security and integration issues. In spite of this, the aerospace industry was stable with a steady demand for their products; Exostar experienced economic risks from its competitors. The year 1999 was a tough year of competition for Chemconnect, who experienced pressures in pricing, loss of export markets, and growing complexity. In 2002, ChemConnect's commodity markets fell due to the post-Enron fears of credit risks.

The Internet enabled construction industry trading partners to access directly other construction-related Web sites. For example, an architect creates economic risks by accessing the site of a lighting fixture manufacturer and obtains the product-specific information needed, thereby bypassing Construction.com's catalog. Pantellos group offered professional services, consulting services, and procurement settlement services. Although it had a business model that does not depend on auctions, it still experienced economic risks.

Farms.com trading partners demanded value-added services. The different kinds of information included real-time, relevant, accurate, and trusted information that led to increased cost. Omnexus faced threats from indirect competitors (i.e., strategic decisions.

Technological Risks

Despite applying Sun's open standards, Covisint experienced technological complexity and administrative disputes leading to integration issues. It was pressured to provide a centralized marketplace for parts auctions and collaboration for its 40,000 trading partners.

Exostar feared that its technological solutions will not be compatible with other browser and operating systems. Furthermore, security, quality, and safety continued to be its critical challenge, as it was a complex industry. Chemconnect experienced integration issues when getting their internal systems to function properly. Construction.com feared the isolation of information assets into separate databases. The lack of integration made it impossible to provide relevant and timely information to their trading partners.

Pantellos group had specific interfaces to handle the transfer of data between its legacy systems using EDI. Although EDI was a reliable system, it was difficult for smaller trading partners to adopt, as it was inflexible and expensive. When eHarvest.com merged with Farms.com in April 2000, it faced challenges in integrating its database. Further, scalability issues arose due to its large databases that, in turn, led to technical components including online auctions, exchanges, farm management solution software, pricing indexes, and user interfaces not promptly delivered. Similarly, Omnexus faced functionality and flexibility issues as its Ariba software had to handle all aspects of the transactions.

Implementation Risks

Covisint had to make a cultural transition from an automotive manufacturing plant to an e-commerce environment. Some of the challenges it faced included:

- How to accelerate business in a technological arena that was unproven.
- How to support the growth of the business through various stages of development by providing different types of services.

It was exposed to fixed costs by creating a traditional business up to a year before those resources were really needed or the technical skills defined. Further, it had to launch a skeleton organization with the risk of being overwhelmed by initial demands.

Exostar had a vast array of managerial and technological challenges, as it was a complex industry. Chemconnect's and Construction.com's implementation risks were derived from using ad hoc software development methods that increased their implementation costs, as it demanded highly specialized and expensive programming skills. Furthermore, the software development processes lacked consistent methods for documenting, saving, and reusing pieces of programming code.

Farms.com faced the emergence of a new generation of Internet-experienced farmers who were demanding more than an online presence. Omnexus' experienced neutrality in the eyes of the suppliers, buyers, and, more importantly, the Securities and Exchange Commission. Its main concern was related to antitrust.

Relational Risks

Covisint experienced relational risks with itssuppliers who faced high operation costs and a lack of strategic vision. Automotive suppliers feared that the lack of security would allow competitors to see their pricing structures and design diagrams. Auto executives acknowledged that it was difficult for many salespeople, engineers, and others to embrace the intangible nature of an e-marketplace. Exostar feared that the e-business exchanges would eventually consolidate with other e-marketplaces. Further, participation required a government certification imposed by the Federal Aviation Administration and the Defense Department.

Chemconnect and Construction.com relational risks were derived from their trading partners who had to subscribe and log onto each service separately and use the company's printed list of products to find specific information. This made searching for products time-consuming, as trading partners could not find the products to meet their specific needs, since the licensing model was complex, and the location-specific information was complicated to use. The major areas of concern for Pantellos Group were credit risks; including liquidity, system security, and operating reliability. Farms.com had to face its big buyers who were forming consortia that threatened their businesses, although the agricultural market was still relatively new.

Omnexus experienced implications of change management in the form of channel conflicts. Its biggest challenge was communicating and demonstrating how its services could be beneficial to its trading partners.

CONTROL MEASURES FOR E-MARKETPLACE RISKS

Controls refer to responses taken by the e-marketplace firms to manage, mitigate, and eliminate their risks. It is part of a facility, including any system, procedure, process, or device that is intended to eliminate hazards, prevent hazardous incidents from occurring, and reduce or mitigate the severity of consequences of any incidents that do occur. We identify four types of controls: economic, technological, implementation, and relationship.

Economic Controls

Covisint developed tools and services that assisted its collaboration and procurement activities by integrating key e-procurement and supply chain functions, thereby saving time and costs on telephone calls and responding to e-mails. Exostar's economic controls were derived from the buying power of its four companies. ChemConnect tried to expand its services by offering new procurement services that aggregated small volume purchases. By doing this, it hoped to achieve repeat transactions, thereby controlling its

economic risks. Construction.com had to use an innovative approach to utilize its intellectual property in multiple market segments and multiple distribution channels.

Farms.com based its Web service development and distribution systems on providing easier Internet-based access to its existing products and services, interconnected product databases, personalized information services, value-added products, and partner applications. Pantellos Group had a menu of supply chain and collaborative applications that provided savings in time and cost. Its exchange moved into an e-procurement system through a strategic alliance with WorldCrest, a procurement service firm giving access to indirect goods and services. This would enable its trading partners to experience economies of scale by not having to spend a lot of time and money to develop its own online procurement processes.

Technological Controls

Covisint's sale of auction services was an evolutionary step toward focusing its strategy on automotive industry operating systems, delivering supplier management portals and data messaging services. Since, Exostar's customers experienced problems running Internet solutions using the old versions of software, they were directed to upgrade their system.
ChemConnect developed its own GXS network to facilitate e-marketplace operations, thereby integrating its key business functions into one compatible system. Construction.com transformed its information service by connecting its isolated data assets and developing new products and services on a Microsoft.NET system that was connected to the Web service architecture, thereby promoting cost-effective data reuse and exchange that, in turn, enabled it to provide customer-defined information to individual subscribers.

These new capabilities provided an integrated database for its projects, products, and people, thereby increasing economic returns from new products and services. Farms.com transferred its software and technology to a focus on supplying trading partners with real-time marketplace and risk management software solutions.

Pantellos Group integrated its system with other technologies, applying Extensible Markup Language (XML). Omneux tried to simplify its business operations by combining disparate sources of data and information, from conflicting technologies and setting up the operating platform.

Implementation Controls

Firms implemented compatible systems in order to manage their implementation risks. Covisint implemented a new format for online trading and relationship building that aimed at collaborative functions. Exostar implemented 87 security requirements, including the highest level of data encryption for both transacting and storage purposes.

ChemConnect implemented a GXS network for its trading partners, who used the Chem eStandards to send and receive documents compatible with EDI. Construction.com formed a partnership with Microsoft to implement both a development platform and a runtime environment that integrates standard construction industry line-of-business applications and Microsoft Office System programs. Pantellos Group formed an agreement with TruSecure® Corporation, a leading security services provider, to offer security, assurance, and certification services. Farms.com employed the latest technology to combine three important components: high quality content, online community, and the possibility of engaging online transactions on one Web site. Omnexus built its model to achieve several supply-side benefits, including integration, increased transaction speed, cost reductions in customer acquisition and retention, automation of the demand chain, real-time inventory and price updating, and alternative purchasing experience for new and old customers.

Relationship Controls

Covisint formed groups of personnel from both consulting and technology backgrounds to resolve relationship issues. Their strategy called for a seamless blend of skills ranging from the consulting to technology spectrum. Exostar implemented an open IT architecture so its trading partners could join without extensive investments in technology. The company also stressed that sensitive information be encrypted according to industry standards. ChemConnect added new settlement services that were expected to reduce relationship risks. Farms.com focused on improving

Table 1. The risk-control framework for e-marketplaces

E-Marketplace Risks	E-Marketplace Controls
Economic Risks	***Economic Controls***
Increased costs of transacting	Focused was on collaborative and procurement services
Pricing pressures	Used the buying power of its founders to create economic returns
Threats from indirect competitors	Expanded its services and provided new procurement services
	Developed greater value added products with less time to market
EDI was expensive and inflexible compared to other Internet based e-commerce applications	Provided easier access to information
	Provided e-procurement services in order to save time and costs
	Specialized in services instead of depending on product sales Provided more value added services and functionality
Technological Risks	***Technological Controls***
Technical complexity	Focused on supplier portals and data messaging services
Incompatible applications and technologies	Upgraded recommended customer configuration
	Provided a stable and reliable IT infrastructure
Lack of integration	Implemented a platform independent technology to integrate databases
Lack of security	Implemented open standards
Failed to deliver technical solutions and services	Implemented their IT system in phases in order to reduce technological risks
Implementation Risks	***Implementation Controls***
Cultural transition from traditional commerce to e-marketplaces	Formed a new service model that focused on collaboration and procurement activities thereby enforcing uniform standards
	Founded a neutral exchange that fulfilled their security requirements.
Lack of uniform standards	Implemented the GXS network to simplify and facilitate IT integration among trading partners.
Lack of consistent operations	
Poor business practices (using adhoc software development methods)	Formed an alliance with a major software development company in order to avoid integration issues and enforcing standard operating procedures.
Management and technical challenges	Formed an agreement with TruSecure® to provide security services in order to enforce quality services.
Lack of technical knowledge and uncertainties	Formed strategic alliances to assist their implementation process in order to prevent technical uncertainties.
Relational Risks	***Relationship Controls***
Imbalance of power	Used outsourcing as a form of collaboration in an untested market
Opportunistic behaviors by trading partners	Added new settlement services in order to reduce risks.
	Changed the focus from a product centric to a customer centric
Concerns with trust due to the intangible nature of the e-marketplace	Formed strategic alliances with third-party application developers and information providers in order to enforce collaboration among trading parties
Searching for products that were time consuming	Provided value added services to their trading partners
	Added more functionality to maintain their e-procurement process
Challenges with change management leading to communication issues	Formed mergers in order to maintain their leadership role
	Provided specific services targeted towards their suppliers in order to sustain long-term relationships
Lack of strategic vision	

customer satisfaction by providing quick and easy access to its information. Pantellos Group allowed for global reach, unlimited operations, and volume capacity, as the Internet presented opportunities for suppliers to expand into new markets. Omnexus recruited skilled personnel and automated its system's functions to capture the demands of its trading partners.

THE RISK-CONTROL FRAMEWORK FOR E-MARKETPLACES

Based on the findings of risks and controls experienced by the seven firms, we developed a risk-control framework for e-marketplace participation (see Table 1). The framework highlights the common key char-

acteristics in terms of risks and controls experienced by each industry type. For example, the study found that automotive suppliers experienced the following risks: security risks when suppliers' competitors got to view proprietary designs transmitted online; standard risks when suppliers purchased expensive and complicated software that was replaced with new systems; implementation risks when the software, online exchanges, and other digital initiatives failed to deliver on their promises; early-adopter risks when a well-intentioned supplier took the plunge but then became a guinea pig or beta test for the rest of the industry; opportunity risks when suppliers spent millions or billions of dollars on software and technology that could have been spent on core competencies such as factory upgrades, wage increases, or other more tangible benefits. Covisint was among the first e-marketplaces to be launched. It had no other models to follow and was restricted by the limitations of the technology. ChemConnect rated its relational risks high due to the Enron case. They had fears of credit risks. This, coupled with the economic downturn, put ChemConnect under pressure to ensure that trading on its exchange would be fair. Technological risks included integration of the e-marketplace with compatible systems, thereby preventing non standardized communication formats.

Based on the key findings of the e-marketplace risks and controls from the seven firms, Table 1 presents the risk-control framework for e-marketplaces.

CONCLUSION

This article identified four types of e-marketplace risks: economic, technological, implementation, and relational. Then it identified four types of control measures to help reduce and manage these risks: economic, technological, implementation, and relationship controls experienced by seven e-marketplace firms across different industries. Based on the key findings of the seven firms, a risk-control framework for e-marketplaces was developed.

E-marketplaces are still formulating standards to meet application requirements that are modular, flexible, adaptable, scalable, robust, and transparent. As emerging and evolving e-technologies offer new opportunities to organizations to conduct businesses globally, it is pertinent to protect their business and customers from fraudulent acts and misuse of the technologies and information. They must ensure that relevant standards incorporate trust-building measures in order to provide globally recognized certificates and credentials, using a wide range of optional encryption standards that are easy to implement, reliable, and user-friendly. The challenge is not the technology. The challenge is the business rules, trust, and basics of e-business. Some folks are having a harder time than others thinking they can share services, content, and business practices, because they feel it is a proprietary business advantage. That is the sort of challenge we face every day.

This article contributes to the security literature as it identified risks and controls those e-marketplaces across different industries experienced. Furthermore, it contributes to practice, as e-marketplace practitioners will be made aware of the kinds of risks and the types of control measures they can enforce for their industry type. Future research should focus on applying this framework with other e-marketplace firms in order to derive a generic framework for e-marketplace participation.

REFERENCES

Bailey, J.P., & Bakos, Y. (1997). An exploratory study of the emerging role of electronic intermediaries. *International Journal of Electronic Commerce, 1*(3), 7-20.

Bakos, Y. (1998). The emerging role of electronic marketplaces on the Internet. *Communications of the ACM, 41*(8), 35-42.

Choudhury, V., Hartzel, K., & Konsynski, B. (1998, December). Uses and consequences of electronic markets: An empirical iInvestigation in the aircraft parts industry. *MIS Quarterly*. 471- 507.

Christiaanse, E., & Markus, L.M. (2002). Business-to-business electronic marketplaces and the structure of channel relationships. *Proceedings of the Twenty-Third International Conference on Information Systems*, 1-9, Barcelona, Spain.

Christiaanse, E., & Markus, L.M. (2003). Participation in collaboration electronic marketplaces. *20th*

Hawaii International Conferences on Systems and Science, Big Island, Hawaii.

Cuny, T., & Richardson, S. (2001). Shopping for the right e-marketplace. *Info Week,* 10-12.

Forrester Research. (1999). Corporate electronic commerce prediction. Retrieved November 23, 1999, from *http://thestandard.net/metrics/display/0,1283,865,00.htm*

Gulledge, T., & Mason, G. (2000). B2B e-marketplaces and small- and medium-sized enterprise. *Manufacturing Information Systems: Proceedings of the Fourth SME International Conference*, 1-7, Parana, Brazil.

Hicks, M. (2001). Survival course: A few independent e-markets bucking the trend thrive. *e-Week*, 37-45.

Kaplan, S., & Sawhney, M.S. (2000). E-hubs: The new B2B marketplaces. *Harvard Business Review. 78*(3), 97-104.

Le, T.T. (2002). Pathways to leadership for business-to-business electronic marketplaces. *Electronic Markets, 12*(2), 112-119.

Lenz, M., Zimmermann, H.-D., & Heitmann, M. (2002). Strategic partnerships and competitiveness of business-to-business e-marketplaces: Preliminary evidence from Europe. *Electronic Markets, 12*(2), 100-111.

Premkumar, G.P. (2003). Perspectives of the e-marketplace by multiple stakeholders. *Communications of the ACM, 46*(12), 279-288.

Reshaur, L.M., & Turner, T.J. (2000). Limiting risky business. *Electronic News (North America), 46*(12), 40-41.

Santos, C., & Perogianni, M. (2001). Electronic marketplaces: Challenges for policymakers, a view by DG Enterprise on the European Commission. *TA-Datenbank-Nachrichten, 4*(10), 30-39.

Segev, A., Gebauer, J., & Farber, F. (1999). Internet-based electronic markets. *Electronic Markets, 9*(3), 138-146.

Sims, I.M., & Standing, C. (2002). Issues in the development of e-marketplaces: A public sector perspective. *Issues and Trends of IT Management in Contemporary Organizations*, 731-735.

Sklar, D. (2001, April). Building trust in an Internet economy. *Strategic Finance*, 22-25.

Vaidyanathan, G., & Devaraj, S. (2003). A five-factor framework for analyzing online risks in e-businesses. *Communications of the ACM, 46*(12), 354-361.

White, A., & Daniel, E.M. (2003). Electronic marketplace-to-marketplace alliances: Emerging trends and strategic rationales. *ACM*, 248-258.

Wise, R., & Morrison, S. (2000, Nov/Dec). Beyond the exchange—The future of B2B. *Harvard Business Review*, 86-96.

KEY TERMS

Control Measures: Responses taken by e-marketplace firms to manage, reduce, mitigate, and eliminate their risks.

Economic Risks: Risks derived from increased transaction costs that lead to reduced financial returns.

E-Marketplace: An interorganizational system through which multiple buyers and sellers interact to accomplish one or more of the market-making activities.

Implementation Risks: Risks derived from poor business practices such as lack of training, lack of uniform standards, quality, and procedures that causes dissatisfaction among trading partners.

Relationship Risks: Risks derived from imbalance of power among trading partners who exercise opportunistic behaviors that, in turn, lead to poor reputations and lack of business continuity.

Risks: Risks can be viewed as a hazard, weakness, uncertain outcome, or opportunity.

Technological Risks: Risks derived from incompatible technologies that cause integration and operation issues.

Road Map to Information Security Management

Lech J. Janczewski
The University of Auckland, New Zealand

Victor Portougal
The University of Auckland, New Zealand

INFORMATION SECURITY ISSUES

Developments in multimedia technology and networking offer organizations new and more effective ways of conducting their businesses. That includes intensification of external contacts. Barriers between different organizations are becoming less visible. The progress gives advantages to competing forces, as well. In the past, an organization was directly exposed to competition only within its own region. Now, due to easy communications, a competitor could be located on the opposite side of the globe, having the ability to access or even disrupt the most sensitive information of a competing company. Hackers and other cyber-criminals are another part of the external threat.

Thus, advantages of using multimedia technology and networking could be accomplished only if data handled by a company are *secure* – available only to the authorised persons (*confidentiality*); represent true values – are not changed during storage, processing or transport (*integrity*); and are available on demand (*availability*). Managing security of information becomes an essential part of running any modern IT system.

This article presents a first-level guidance for how to approach this problem.

The most widely known document on information security is an annual *Computer Crime and Security Survey (CCSS)*, conducted by San Francisco's Computer Security Institute in cooperation with the Federal Bureau of Investigation (FBI) (CSI, 2003). It is based on responses from more than 500 professionals representing all types and sizes of organizations, from huge international corporations to small businesses, from nationwide government agencies to small community centres. The message the survey conveys is frightening:

- The total annual losses reported in the 2003 survey were more than $200 million.
- As in prior years, theft of proprietary information caused the greatest financial loss (more than $70 million was lost, with the average reported loss being approximately $2.7 million).
- In a shift from previous years, the second most expensive computer crime among survey respondents was denial of service, with a cost of $65 million.
- Losses reported for financial fraud were significantly lower, at $10 million.
- As in previous years, virus incidents (82%) and insider abuse of network access (80%) were the most cited forms of attack or abuse.

The report is covering only a very small part of the United States' (U.S.) economy; real nationwide losses could be several magnitudes higher. Surveys of a similar nature are conducted in many other countries, like Australia (AusCERT, 2003). These surveys brought similar results. It is not a surprise, as the whole globe is becoming a wired village, and computer technology is the same all over the world.

These alarming facts are now a major worry of the business community. This is reflected in surveys asking organization executives what their main points of concern are and which activities they consider the most important. Two decades ago, information security issues were nonexistent in these surveys. They had appeared on the "Top 10" list around the early 1990s, and they are gradually making their way towards the top. Bombarded by the flood of warnings about possible damages from the misuses of information technology, managers switched to investing in security measures. However, these investments are done quite reluctantly. The nature of

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

threats is still mysterious to non-specialists, and one of the most common statements is: "Why should I invest in information security when we did not register any abuses or attacks?"

Unfortunately, unlike bank robbery, many attacks against computers are difficult to notice and thus impossible on which to launch an investigation. The classical example is hacking: attempts to gain unauthorised access to computer resources. If the hacker was either unable to break into the system or did not change any records, then such an attempt would remain unknown if the installation did not have any hacker-detecting tools. The possible consequences could emerge much later and may not necessarily point to a particular hacker attack. Of course, ordinary information systems with highly sensitive information need protection from hackers. Intrusion detection methods have been developing over the past half-decade, largely in response to corporate and government break-ins (Durst, Champion, Witten, Miller & Spagnuolo, 1999). In many cases, when appropriate detection tools had been installed, the information technology managers were terrified to learn about the extent of their system abuses.

Two essential strategies exist for protection of network infrastructures. One is a "terminal defence" initiative undertaken by the owners of individual nodes in a network to protect their individual nodes from persistent, well-supported intrusion. The other strategy is a "collective action" that involves groups of owners, industry groups, government groups and so forth who audit the collective system operation and exchange information to detect patterns of distributed attacks. Collective action can also involve redundant capacity across the collective system and the ability to reallocate a system load or to ration diminished system capacity. Both strategies can also involve preventive measures, such as research and development to improve the state of the art in system security or the exchange of threat and countermeasure information (Lukasik, Greenberg & Goodman, 1998).

Intrusion detection attempts to discover attacks, preferably while they are in progress or at least before much damage has been done. Automation of intrusion detection is typically premised on automated definition of misuse instances. This automation requires pattern recognition techniques across large databases of historical data. Methods for data mining clearly have contributed to making such intrusion detection feasible (Bass, 2000; Zhu, Premkumar, Zhang & Chu, 2001). These approaches have been growing in sophistication, and include expert systems, keystroke monitoring, state transition analysis, pattern matching and protocol analysis (Biermann, Cloete & Venter, 2001; Graham, 2001).

But intrusion detection approaches thus far remain a probabilistic enterprise, with less than 100% chance of detecting all types of intrusion. Indeed, the race between intruder technology and intrusion detection will likely remain a closely run contest. New tools make attacks undetectable. Intrusion detection tools are necessary, but not sufficient for the high-stakes information resources subject to attacks.

The typical approach to information security is labeled *piecemeal approach*. Many information security tools are well known, like firewalls or virus scanners. Under the piecemeal approach, the user sees danger of a specific threat, identifies tool(s) to reduce such a threat and implements this tool. Such approach may work, but it would not necessarily render the optimal solution from the overall perspective of a business organization.

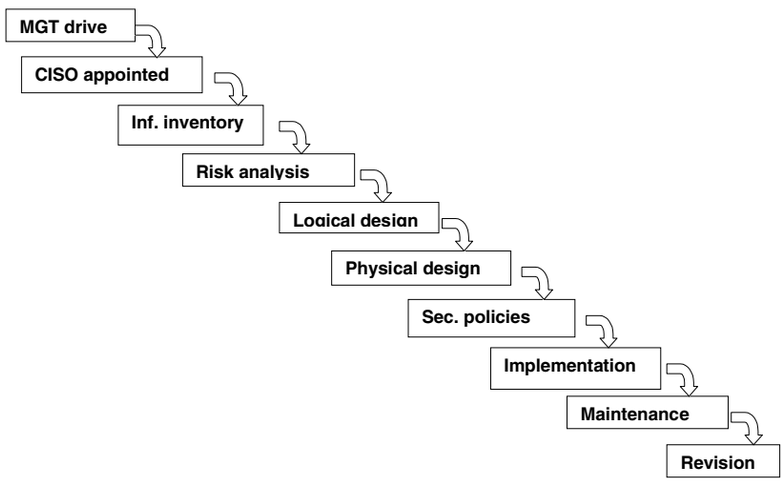
INFORMATION SECURITY MANAGEMENT

System approach is a top-down methodology of developing an information security system recommended in literature. It is based on an IBM-developed methodology of investing in information technology called Business System Planning. The following process is a modification of that methodology to the needs of information security. The basic 10 steps of the methodology are presented in Figure 1.

Step 1: Managerial Drive

The building of a sound security system should be initiated, endorsed, supported and controlled by the top management. The IT personnel may have a very sound understanding of the information security processes or of their importance, but without top management's understanding and active support, introduction of an effective security system is impossible.

Figure 1. Basic steps in the managerial approach to information security



Hence, the first step of the system approach is to convince management that information security issues must be taken seriously. Many methods of raising such awareness could be implemented. One of the most effective is shock therapy, based on a live demonstration to top management of the vulnerabilities of the information technology in their organization.

It is a recognized fact that many business organizations are running wireless Local Area Networks (LAN) without activating any security measures. This means that anybody equipped with a laptop and a wireless LAN adapter is able to switch into a company network and monitor all the interior traffic. Presenting top management with some leaked documents usually jolts them into taking security issues seriously.

However, such demonstrations should be applied very carefully, as in several countries laws prosecute unauthorised access to computer networks.

Step 2: Appointment of Chief Information Security Officer (CISO)

The naming of the position is not important. What is important is that within every organization, huge or small, there should be a person appointed to be responsible for information security issues. Depending on the size and type of a company, it might be a part-time or a full-time position, or the person may even head a team. For instance, there will be more information security officers in a bank than in a chain of retail shops. In an average business organization relying heavily on information technology, there should be one security officer per 500 employees. There are a number of issues related to appointment of the CISO.

The first issue is the required qualification for such a job. Having good programming and system analysis skills are a good enhancement for a candidate, but preferences should be given to holders of certificates confirming their information security knowledge. These qualifications are currently awarded by such organizations like Cisco, Symantec and Microsoft, or specialised bodies like International Information Systems Security Certification Consortium (ISC)[2] or Computing Technology Industry Association (Security+). Many universities around the world are successfully running courses of studies in information security, too.

The second issue is the placement of the CISO in the organisational hierarchy. The CISO job is rather above the development and use of information technology. Therefore, the CISO should not be a member of the information technology department or division. The optimal solution is to make the

CISO report directly to the CEO of the organization or his deputy (but not the deputy responsible for information technology). In some organizations, duties of the CISO are combined with duties of an officer responsible for general security. In an environment where general security of the establishment is central (like some government or military units), such a solution is reasonable.

In large organizations, setting up of a Security Advisory Committee, controlling the processes of development, implementation and running of the security system, should follow the appointment of the CISO. The committee should include leaders of all major divisions of the company. Such a committee is not aimed on actual development of any policies or programmes, but rather on ensuring that any such decision would work to the benefit of every division of the company.

Figure 2 illustrates organization structure for information security functions.

Step 3: Information Inventory

In this step, the information has to be collected about all data important for the company's operation, both current and planned. The data format is not important (traditional documents or electronic files). The important information is:

- The value of the data for the company.
- The processes: generation, processing, storage and possibly disposal.
- The owner of the data and the handler(s).

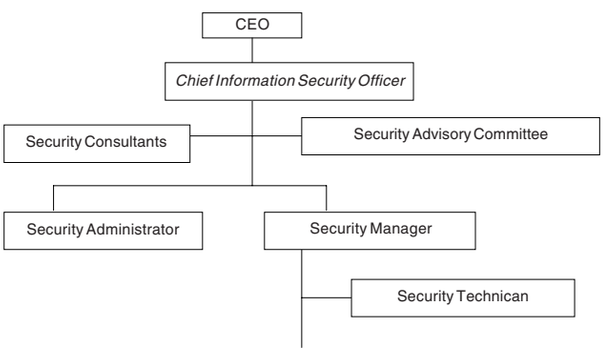
After this, rules need to be set up regarding handling of the data. Each item of data needs to be assessed, and the following characteristics should be assigned:

- Security category.
- Expiry date for security level.
- Ownership of the data.

At the same time, rules related to the security clearances of company employees need to be set up. These rules determine which documents (in terms of their security level) a person is allowed to handle. These procedures need to be established prior to the development of any security systems. The resulting security system objective is simply to enforce these procedures.

The weak point of this methodology is the issue of assigning security clearances to an individual. In a typical business environment, this procedure is based on a position of a given person within the hierarchy of an organisation. The general principle is that "the higher you are within the company hierarchy the highest security clearance you must have." Such an approach clearly incurs significant problems. On the one extreme, a person might have security clearance too high for the job, which increases the total cost of the security system. The higher the security clearance, the higher the cost (for example, for security training). On the other extreme, a person with a security clearance too low for the job must obtain temporary authority for accessing specific documents. Such a procedure

Figure 2. Structure of information security functions



could be costly and time consuming, and decrease the efficiency of operations. Portougal and Janczewski (1998) demonstrated in detail the consequences of the described approach in complex hierarchical structures.

A competing and more logical idea is to apply the "need-to-know" principle. Unfortunately, in its traditional formulation, this principle does not give adequate guidance to the management of how to set up security clearances for each member of the staff. Amoroso (1994) describes the "principle of least privilege." The recommended application is based on subdividing the information system in certain data domains. Data domains may contain secret or confidential information. Users have privileges (or rights to access) to perform operations for which they have a legitimate need. "Legitimate need" for a privilege is generally based on job function (or a role). If a privilege includes access to a domain with confidential data, then the user is assigned a corresponding security clearance. The main flaw of this approach is that a user has access to the whole domain even if that user might not need a major part of it. Thus, the assigned security clearance may be excessive. A similar problem arises regarding the security category of an object. A particular document (domain) could be labeled "confidential" or "top secret" even if it contains a single element of confidential (top secret) information.

Portougal and Janczewski (2000) suggested another realisation of the "need-to-know" principle. Their method is based on Data Access Statements (DASs) defined for all employees as part of their job description. DASs list all data elements needed by an employee to perform duties effectively. Thus, they shifted the assignment of security clearance from the domain level to the element level. This approach solves the difficult problem of defining individual security clearances. In addition, it connects this problem to more general problems of the security of the organisation as a whole, to the problem of security cost and cost optimisation.

Step 4: Risk Analysis

Risk analysis is the continuation of the analysis performed at the previous step. Its objective is to define to what extent particular data could be the subject of unauthorised access or alterations. In essence, two questions are asked about each data element:

- What is the probability that the information would be subject to unauthorised read or write operation?
- What losses would result from compromising such data?

The questions are simple, but finding accurate answers to these questions usually is almost impossible, because:

- The cost and duration to collect such probabilities may be so huge that the job will not be acceptable for management.
- Attacks never happened, but they may happen in the future, so there is no reliable loss of information.

Probability theory is the most commonly used theoretical base in the process of risk analysis for information systems security. When using probability theory models, certain assumptions are made about distributions of the intruder population and of the intrusion events. In commercial settings, for example, the intruder population might be assumed to be distributed normally, while the intrusion events have a Poisson distribution (Baskerville & Portougal, 2000). These assumptions are not valid in security settings where intruders are very highly skilled and intrusion events are persistent; that is, attacks based on infinite intrusion resources. Such security settings include, for example, those where national information infrastructures become enveloped in a battle space subject to information warfare attacks.

Baskerville and Portougal (2003) introduced an alternative quantitative approach to information security evaluation suitable for information resources that are potential targets of intensive professional attacks. This approach originates in the recognition that the safety of all information resources in an organization is only *an opinion* of officials responsible for information security. The set of security measures introduced to protect information resources depend on the experience of security personnel, their perceptions and their attitude toward risk. This approach is often called *Delphi methodology*.

To ease the task of performing a risk analysis, a number of methodologies have been prepared, and we strongly recommend getting acquainted with some of them.

Step 5: Logical Design

The basic decisions regarding the components of information security systems are determined during this phase of the security system building. These components may be subdivided into four groups:

- **Security hardware:** Firewalls, hardware telecommunication monitors, physical perimeter protection and so forth.
- **Security software:** Virus scanners, intrusion detectors, encryption facilities and so forth.
- **Security organization:** Development of various policies related to secure processing of information; here also may be included the procedures of setting security categories of data.
- **Personnel:** Development of security policies related to hiring new staff, security training and discharging.

This step is not aimed at developing the final format of any of these documents; but rather, for establishing a framework for each of them following the company's overall security requirements. As a result, it is assured to a high degree that any detailed solution of the security system would follow the overall objectives set up by top management.

Step 6: Physical Design

Here the logical design guidelines are put in practice. Decisions are made on the source of each physical component of the security system (development in-house, development on request or a shrink/wrap product).

Physical design should follow all the operations described in previous stages. Only then may the organisation be sure that the resulting security system would be effective. Unfortunately, for many companies this is the starting point of their security system design. These companies, too frequently, have counterproductive results.

A very important point needs to be mentioned: Change the default settings of the installed components immediately after installation. Default settings are values of the main parameters set up by the manufacturer of the product to allow easy installation. For instance, firewall parameters could be set up in such a way that the firewall would be totally transparent for traffic from both directions, and a password of the system could be set simply to *password*. Surprisingly, many buyers of security products, for some reason, are not changing these values. The values are commonly known and could be used by hostile people to compromise the company's resources.

Step 7: Security Policies

Security policies determine procedures related to:

- hiring new staff
- security training of the staff
- secure handling of company data
- discharge of employees.

A document that specifies all the fundamental security duties of every staff of an organization must be developed. This document frequently is called the *Information Security Policy* (ISP). The document is usually presented to a person upon commencement of employment with a request to read, sign and follow. The document would not outline all detailed procedures, but would define all duties related to the security issues each employee should follow.

Developing an effective ISP is not an easy task. It is recommended to use some guidelines, like the international standard ISO 17799 on managing information security, as a starter. Then a quicker result may be achieved without decreasing the quality of ISP.

Step 8: Implementation

At this point, all the developed procedures and solutions are put to practical use. There are no special rules related to this stage. All good practices of systems implementation should be tried. Perhaps one point is extremely important: It is the end user's involvement. The attitude of end users, dealing every day with the security system, is crucial. To avoid hostile attitude of users, their cooperation needs to be gained. The best way to achieve this is

to engage them in the development process on all stages. By being involved, the users might (hopefully) develop a positive attitude toward the offered solutions, then treat them as a regular part of their duties.

Step 9: Maintenance

Nothing is more annoying as a component of a system that is supposed to work, but does not work properly. Initial trust put on this component should not be jeopardized by sloppy maintenance. The aim of maintenance procedures is that all parts of the security system work according to their design.

It is recommended to automate system maintenance. For instance, when installing virus scanners or intrusion detectors, it is worth arranging automated delivery and installation of the available updates. Of course, it has to be done in a secure way; that is, that the updates are real updates, generated and transported from the original source, and not compromised on the way. When the maintenance process is outsourced, the company must be sure that only the authorised party does the remote maintenance.

Step 10: Revision

Revision introduces changes to the technology, user requirements or environment. Revisions should be done on a regular basis, though it is very important to provide an adequate reaction to such rapidly changing conditions, which otherwise could put the whole system in disarray.

FUTURE TRENDS

One way or another, information security is a race between hackers, cyber-terrorists, cyber-criminals and owners of IT to provide proper functioning of computers and networks. Each day, mass media and specialized sources bring information about new attempts to violate the integrity of IT and tools to confine these threats.

A majority of the existing procedures have a retroactive nature – they attempt to fix security holes that have been found. To gain in this race, we need to adopt a proactive approach: Predict possible attacks and plan for them. The 10-step procedure of planning a security system, described in the text, is perhaps an example of such a trend and should be continued.

To accomplish secure computing, management must seek the advice and help of an experienced security specialist. But whoever that specialist is, the major goal of the whole exercise must be the increased success of the whole organization. A company may enjoy benefits of using multimedia technology and networking facilities to their advantage only if it has an information security system of high quality.

REFERENCES

Amoroso, E. (1994). Fundamentals of computer security technology. Prentice Hall.

AusCERT. (2003). Australian computer crime and security survey. Retrieved from *www.auscert.org.au*

Baskerville, R., & Portougal, V. (2000). A framework for evaluation of security provided by firewalls (working paper). Atlanta: Georgia State University.

Bass, T. (2000). Intrusion detection systems and multisensor data fusion. *Communications of the ACM, 43*(4), 99-105.

Biermann, E., Cloete, E., & Venter, L.M. (2001). A comparison of intrusion detection systems. *Computers & Security, 20*(8), 676-683.

CSI. (2003). Computer crime and security survey. Retrieved from *www.gocsi.com*

Durst, R., Champion, T., Witten, B., Miller, E., & Spagnuolo, L. (1999). Testing and evaluating computer intrusion detection systems. *Communications of the ACM, 42*(7), 53-61.

Graham, R. (2001). NIDS - Pattern search vs. protocol decode. *Computers & Security, 20*(1), 37-41.

Lukasik, S.J., Greenberg, L.T., & Goodman, S.E. (1998). Protecting an invaluable and ever-widening infrastructure. *Communications of the ACM, 41*(6), 11-16.

Portougal, V., & Janczewski, L. (1998). Industrial information-weight security models. *Information Management & Computer Security, 6*(5).

Portougal V., & Janczewski, L. (2000). "Need-to-know" principle and fuzzy security clearances modelling. *Information Management and Computer Security, 8*(5), 210-217.

Zhu, D., Premkumar, G., Zhang, X., & Chu, C.-H. (2001). Data mining for network intrusion detection: A comparison of alternative methods. *Decision Sciences, 32*(4), 635-660.

KEY TERMS

Business System Planning (BSP): IBM's developed methodology of investing in information technology. An example of a *system approach* methodology of developing an information system.

Chief Information Security Officer: Employee of an organization who is top authority in relation to information security issues.

Collective Action: An initiative, undertaken by groups of owners, industry groups, government groups and so forth, who audit the collective system operation and exchange information to detect patterns of distributed attacks.

Information Security: Domain of knowledge dealing with issues of preserving confidentiality, integrity and availability of information.

Information Security Policy: A document that outlines the basic rules of safe processing and dissemination of information.

Intrusion Detection: Attempts to discover attacks while they are in progress, or at least discover them before much damage has been done.

ISO 17799: International standard describing managing information security processes.

Need-to-Know Access Policy: Security access policy based on supplying to individual employees only information necessary to perform their duties.

Piecemeal Sesign: A method of designing a system in which each component of a system is developed independently.

Principle of Least Privilege Access Policy: Equivalent of "need-to-know" security policy related to the role-based security access model.

Risk Analysis: To what extent particular data could be a subject of unauthorised access or alterations.

Security Category: Limitation of circulation imposed on a document or a file.

Security Clearance: A set of privileges individually granted to an employee related to dealing with confidential information.

Terminal Defence: An initiative, undertaken by the owners of individual nodes in a network, to protect their individual nodes from persistent, well-supported intrusion.

Security Laboratory Design and Implementation

Linda V. Knight
DePaul University, USA

Jean-Philippe P. Labruyere
DePaul University, USA

INTRODUCTION

Security laboratories provide controlled environments that simulate enterprises' infrastructures. Such laboratories allow technical professionals to test the effectiveness of different hardware, software, and network configurations in warding off attacks, as well as to experiment with and learn about various security devices, tools, and attack methods in a controlled manner that insures benign consequences. These laboratories typically include an extensive and sometimes complex networking environment.

This paper identifies the critical issues that make the design and implementation of a simulation environment difficult, and provides ways to address these concerns through a checklist of nine critical security-lab design features. Design and development principles and technical and engineering requirements proposed here theoretically can be of use to businesses or universities seeking to build a security laboratory. They can also provide a useful checklist for managers charged with the IT function to use when discussing their security laboratory with their lab's technical designers and support staff.

Historical Perspective

As organizations depend more heavily upon their information resources, and these resources are more commonly attacked, security laboratories become increasingly important. The number of attacks reported to Carnegie Mellon University's CERT Coordination Center (CERT, 2004) grew from 6 in 1988 (the year it was established) to 21,756 in 2000 and 137,529 in 2003. By 2004, automated attacks had become so prevalent that CERT stopped publishing the number of incidents. Such attacks are costly. According to the 2004 FBI and CSI survey (Gordon, Loeb, Lucyshyn, & Richardson, 2004), the 269 respondents who provided costs estimates on the damages estimated that losses reached $14,496,560 in 2004. Yet, the CSO Magazine, U.S. Secret Service, and CERT/CC 2004 E-Crime Watch Survey (2004) found that 32.4% of their respondents did not track monetary losses due to electronic crimes or system intrusions. According to the U.S. Secret Service special agent in charge of the Criminal Investigative Division, "Many companies still seem unwilling to report e-crime for fear of damaging their reputation" (CSO et al.).

BACKGROUND

The most obvious goal of the security-laboratory environment is to provide a suitable setting for experimentation with computer and network security. Such a laboratory can be used to assess the effectiveness of different configurations against security attacks, as well as to allow laboratory users to experiment with and learn about various tools and attack methods. The difficult question is how to design, deploy, and maintain such a nonproduction or laboratory environment. Key issues revolve around how to provide full functionality without allowing the laboratory to be misused, threatening the security of its parent organization or of outside entities.

Security laboratories may be broadly classified into two types: enterprise and educational. For business enterprises, the security laboratory should mimic the organization's security infrastructure production environment. The lab generally should replicate the organization's core security set and configurations while providing access to data that is fundamentally the same as production data, but without the vulner-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

ability that using actual production data would incur. Within an educational environment, the lab should be designed to follow either the most common or the best-practice recommendations for enterprise security. Such an educational lab is particularly likely to be set up to allow experimentation with a variety of configurations.

Despite growing interest in computer and network security, little research centers on the design of security laboratories for business enterprises. However, several papers do address various aspects of designing security laboratories for university students. Mayo and Kearns (1999, p.165) described a lab where "…students are given complete (root) control of systems with essentially unrestricted access to the Internet." This was accomplished by insuring that clients within the network appear as outside systems, lacking the ability to interact directly with departmental systems. A guiding principle for this design was that students be able to do no more damage than they might from their dorm room. In a related work, Hill, Carver, Humphries, and Pooch (2001) described implementing an isolated laboratory where students in a specific class were divided into two groups: one group with the goal of protecting its computers, and one group with the goal of compromising the other group's computers. A similar situation was detailed by Wagner and Wudi (2004) when they described using a closed network for cyberwar exercises. Matei (2003) offered extensive advice and resources for those wishing to develop a lab-based course on Internet security. This lab also was isolated, with the exception of specific controlled connections to the department's server. In yet another work related to educational security laboratories (Frank, Mason, Micco, Montante, & Rossman, 2003), a five-member panel who had attended a National Science Foundation (NSF) sponsored cybersecurity workshop shared their thoughts on how they applied what they learned to their courses. Themes that emerged in the panel discussion included moral and ethical considerations, the need to isolate laboratory functions, and the need to formally assess risk. These themes were further developed in work by Labruyere and Knight (2004) that is believed to be the first to center upon the design of both enterprise and educational security laboratories. Key principles from this work are incorporated throughout this paper.

CRITICAL ISSUES

The greatest challenges involved in implementing and supporting the security-laboratory environment are, for the most part, the result of seemingly conflicting functional requirements. In particular, the lab must allow the implementation and utilization of dangerous tools while protecting the production environment and Internet-accessible host from such tools. The lab hosts must have access to outside resources for downloading updates, patches, or documentation, and yet the lab must be protected from outside-initiated attacks. Strict logging of all activities must be implemented for control purposes, but the privacy of the lab user must be maintained. The lab must be able to be reinitialized relatively quickly to a stable and secured state, yet support and maintenance resources are likely to be scarce. Finally, the lab must closely mimic the production environment, but no live data must be present and the lab must be set up in a fashion that will not give an intruder useful information concerning the actual production setup and infrastructure.

DESIGNING A SECURITY LABORATORY

Conflicting functional requirements can be addressed by implementing a combination of nine critical technical design features, as described in the text that follows.

Implement Strict Activity Logging

A strict, auditable system is required to control access to laboratory resources. A copy of all activities must be kept on a real-time basis and logged to a repository that is not directly accessible from the lab environment. All communications between the lab devices and the logging facility should be done via out-of-band connections, that is, connections that are not used by the lab or production facilities and that are protected from disruptions and attacks. When logging activity, actual data payloads may be kept or discarded, depending on the organization's legal and ethical requirements. The logging system must include the sending of null-message heartbeats

to alert the lab administrator when a resource cannot perform the real-time logging.

Just as in a production environment, the privacy of the security-laboratory user must be preserved. Unfortunately, this privacy requirement conflicts with the need for an audit trail with strict logging of all activities and accountability for all actions taken. Each organization must find its own balance between privacy and logging requirements by taking into account its own unique legal, ethical, and regulatory requirements and policies. Solutions may be grounded in regulatory requirements, the organization's employee manual, or its acceptable-use policy (AUP).

Although it is impossible to give a generalized solution to the logging-vs.-privacy dilemma, at a minimum, lab users must receive clear warnings that all activities are monitored and logged. This should be done by making the lab user sign a laboratory acceptable-use policy, as well as by displaying warning banners and dialog boxes for all devices accessed by the lab user.

Implement Laboratory-Access Control

Typically, the security-lab environment will use the same Internet connection as the production environment. Access controls must be implemented to ensure the lab resources cannot access the production environment. These access controls could include authentication systems based on the IEEE 802.1x standard (2001), Radius (Rigney, 1997), TACACS (Terminal Access-Controller Access-Control System; Finseth, 1993), or Kerberos (Kohl & Neuman, 1993) protocols.

Enable Restriction on Outbound Traffic Types

Lab hosts must have access to outside resources for downloading updates, patches, or documentation. At the same time, outbound traffic that is malicious or nonauthorized must be prevented. This can be achieved by setting up strict restrictions on the type of traffic and destinations allowed. The remote logging of all activities described earlier can ensure that such controls are in place, functional, and not bypassed. Even organizations that do not routinely store data payloads may wish to do so for outbound traffic. The feasibility of keeping such copies will be determined by the amount of traffic generated, the capacity of the logging facility, and privacy requirements.

Enable Bandwidth Limitations on Outbound Traffic

Very often, a security-laboratory environment is connected to the main Internet link of an organization. That bandwidth is likely to also transport production or mission-critical traffic along side the security-lab traffic. In the case of a large organization, that Internet link may have a high bandwidth capacity. However, the amount of bandwidth allowed to leave a security laboratory must be limited in order to thwart the usage of the lab for denial-of-service (DOS) attacks. In DOS attacks (Douligeris & Mitrokotsa, 2004; Hussain, Heidemann, & Papadopoulos, 2003), a victim site is overwhelmed with high traffic levels. The best method for preventing the use of a security lab to launch such threats is to introduce a restsricted traffic-bandwidth policy or an artificial bottleneck that will limit traffic levels leaving the lab environment. In addition, limiting security-laboratory bandwidth prevents the lab from denying the organization's regular production traffic adequate access to the Internet.

Implement an Efficient Configuration Management and Restoration System

Since the security-laboratory environment will be changed through experiments with alternate setups and test configurations, the security-laboratory administrator must be able to restore the lab in a fast and secure fashion to a known state. That known state or baseline will change dynamically with new patches and frequent configuration changes. Thus, an efficient system must be set up to manage changes in the baseline and perform restoration.

Ban all Production Data from the Security Laboratory

Such a ban may not be particularly meaningful for university-research or student security laboratories; however, it is critical for an enterprise labora-

tory to avoid all risk that production data might be compromised. This can be accomplished by banning all production data from the security lab, and replacing all data in databases or applications with randomized strings and identifiers.

Hide Information about the Production Environment

The security-lab environment, set up to mimic the production environment, must not give a lab user (or a successful intruder) any useful information on the actual configuration or structure of the production environment. This can be accomplished by using a different naming space, implementing a different IP (Internet protocol) addressing scheme, and most importantly, assuring no production document or data are present in the lab environment.

Implement only the Minimal Software Needed

In an enterprise environment, the security laboratory should only implement the minimal software needed to address functional requirements. Adding powerful security tools that are not critical to the lab's charter poses an unnecessary risk. For example, a front-end logging system might be implemented, but the complete functional application should not be made available if it is not part of what is being tested. In a university environment, more than minimal software may be needed to provide a rich educational or research environment. In this case, the institution's managers or administrators should be called upon to make a conscious decision, balancing the added risk against the added educational benefits.

Promote the Ethical Use of Information-Security Resources

An organization that deploys a security laboratory has to take into account many legal and ethical considerations that often are not present in a standard production environment. Since the security lab could be used to perform attacks and might allow a user to gain expertise and skills that could be used later for malicious purposes, legal and ethical considerations come into play. Areas to be taken into account include organizational liability for providing the tools and infrastructure used in an attack, how an organization can preserve the privacy of its lab users while enforcing accountability for their actions, and how the organization might promote the ethical use of security auditing tools. The answers to such legal and ethical questions require a proactive, multifaceted study that includes an organization's human-resource and legal departments. This type of broad study requires considerable time, and can be particularly problematic for smaller organizations without extensive in-house resources readily available. However, it is critical that legal and ethical issues are studied before a security lab is designed and deployed.

When an organization skips the step of requiring a careful up-front analysis of the human and ethical factors by those most skilled in these areas, it is likely to end up having its policies determined by a group that is relatively untrained in human, ethical, and legal considerations: the technical staff that design and build its security laboratory. Although the technical staff can be an important resource in quantifying the amount of risk involved in including various features in the lab, neither risk analysis nor legal and ethical decisions should be left to their discretion. Key decisions in these areas should be made by professional managers and administrators trained in such considerations.

To facilitate the ethical use of information-security resources, an organization may implement mandatory classroom or self-paced training, and/or require all users to sign a code-of-conduct agreement before being granted access. In any case, an individual should be designated as the key person responsible for promoting the ethical use of the laboratory.

FUTURE TRENDS

Proving the soundness of any approach to designing a security laboratory, or even of any one deployed laboratory environment, is in fact not possible. One can demonstrate that a given deployed environment failed by mitigating its integrity. However, the only evidence that such an environment achieves its goal

comes from verifying over time that it has not been compromised. Such evidence cannot be considered proof of the soundness either of the laboratory or of the methodology used to design it. Furthermore, the compromise of an established laboratory environment does not necessarily mean that the methodology followed for its design is flawed. While the compromise may be caused by a defect in establishing the functional requirements, it is even more likely to be caused by a defect in implementation and configuration. Thus, as is often the case in the security domain, the principles described here cannot be proven. However evidence of their soundness and of their weaknesses can be expected to become clear over time as they are employed. The design considerations documented here can be expected to develop further over time as such evidence emerges, and as new technologies and external threats continue to emerge.

CONCLUSION

The implementation of a security-laboratory environment provides many benefits to both business enterprises and universities. Chief among these are the ability of lab users to increase their skills by experimenting with the methods and tools typically used by intruders, and the ability of the lab to be used to test a configuration or a system for security weaknesses before production deployment. The implementation of a security lab does, however, introduce complex threats, and many design aspects must be closely considered. The major areas of concern to address are how malicious activities can be prevented from originating in the security-lab environment; how the lab environment can be protected from attacks and from being compromised; how the privacy of lab users can be maintained while implementing the necessary logging and auditing system to enforce accountability; and what tools and methods can be used to facilitate simple, trouble-free management of the lab environment.

The answers to these concerns include technical solutions, policy decisions, and procedural solutions. While the specific solutions will vary with the organization, some critical principles form the basis for the design of any security lab: Implement access control and strict activity logging; restrict outbound traffic types and limit outbound traffic; implement an efficient configuration-management and -restoration system; ban production data and hide information about the production environment; implement only the minimum software needed; and promote the ethical use of resources through policies, procedures, and the education of lab users. These principles can be used by enterprises seeking to design and develop a security laboratory. They can also provide a checklist of critical considerations for managers to use when discussing their organization's security laboratory with their technical support staff.

REFERENCES

CERT. (2004). *CERT/CC statistics 1988-2004.* Retrieved August 3, 2004, from *http://www.cert.org/stats/*

CSO Magazine, U.S. Secret Service, & CERT Coordination Center. (2004). *2004 e crime watch survey.* Retrieved August 3, 2004, from *http://www.csoonline.com/releases/052004129_release.html*

Douligeris, C., & Mitrokotsa, A. (2004). DDoS attacks and defense mechanisms: Classification and state-of-the-art. *Computer Networks: The International Journal of Computer and Telecommunications Networking, 44*(5), 643-666.

Finseth, C. (1993). *IETF request for comments (RFC) 1492: An access control protocol, sometimes called TACACS.* Retrieved August 6, 2004, from *http://www.ietf.org/rfc/rfc1492.txt*

Frank, C., Mason, S., Micco, M., Montante, R., & Rossman, H. (2003). Panel discussion: Laboratories for a computer security course. *The Journal of Computing in Small Colleges, 18*(3), 108-113.

Gordon, L. A., Loeb, M. P., Lucyshyn, W., & Richardson, R. (2004). *CSI/FBI computer crime and security survey.* Retrieved August 6, 2004, from *http://www.gocsi.com/forms/fbi/pdf.jhtml*

Hill, J. M. D., Carver, C. A., Jr., Humphries, J. W., & Pooch, U. W. (2001). Using an isolated network laboratory to teach advanced networks and security.

Proceedings of the 32nd SIGCSE Technical Symposium on Computer Science Education, 36-40.

Hussain, A., Heidemann, J., & Papadopoulos, C. (2003). A framework for classifying denial of service attacks. *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, 99-110.

IEEE 802.1X-2001. (2001). *IEEE standards for local and metropolitan area networks: Port-based network access control.* Retrieved August 6, 2004, from *http://standards.ieee.org/getieee802/download/802.1X-2001.pdf*

Kohl, J., & Neuman, C. (1993). *IETF request for comments (RFC) 1510: The Kerberos network authentication service* (V5). Retrieved August 6, 2004, from *http://www.ietf.org/rfc/rfc1510.txt*

Labruyere, J. P., & Knight, L. V. (2004). Designing a controlled environment for the simulation of an enterprise security infrastructure. *Innovations through Technology: Proceedings of the 2004 Information Resources Management Association Conference*, 29-32.

Matei, P. (2003). A laboratory-based course on Internet security. *Proceedings of the 34th SIGCSE Technical Symposium on Computer Science Education*, 252-256.

Mayo, J., & Kearns, P. (1999). A secure unrestricted advanced systems laboratory. *Proceedings of the 30th SIGSCE Technical Symposium on Computer Science Education*, 165-169.

Rigney, C. (1997). *IETF request for comments (RFC) 2138: Remote authentication dial in user service (RADIUS).* Retrieved August 6, 2004, from *http://www.ietf.org/rfc/rfc2138.txt*

Wagner, P. J., & Wudi, J. M. (2004). Designing and implementing a cyberwar laboratory exercise for a computer security course. *Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education*, 402-406.

This paper builds upon and includes excerpts from a paper by the same authors presented at the 2004 Information Resources Management Association conference, which is listed in the references above.

KEY TERMS

Acceptable-Use Policy (AUP): A written policy document that defines what activities are appropriate and inappropriate for a user of a particular resource. A document indicating the understanding and acceptance of an AUP is often required to be formally signed by a user before he or she gains access to the resource.

Access Control: Methods used to determine if requests to use a system, network, application, or resource should be granted or denied.

Activity Logging: Electronic record keeping of such system or network actions as applications accessed, commands executed, files accessed, and traffic generated from a system.

Authentication System: System used to verify the identity of an entity (user, application, host, system, and device) that is attempting to participate in a computing environment.

Denial-of-Service (DOS) Attacks: A type of computer-system security attack where an opponent prevents legitimate users from accessing a service or a resource, typically by overloading that resource with fabricated requests.

IEEE 802.1x Standard: A standard that defines port-based network-access control. It is used to provide authenticated network access for local-area networks and, using some extensions, for wireless networks.

Kerberos: A type of authentication, developed at MIT, where a Kerberos server issues a session ticket when a valid user requests services. The user's log-on password is not sent through the network.

Radius: An authentication method that allows remote users to access a server or a network or computing resource.

Security Laboratory: A controlled environment used by technical professionals to test the effectiveness of different hardware, software, and network

configurations in warding off attacks, as well as to experiment with and learn about various security devices, tools, and attack methods.

TACACS (Terminal Access-Controller Access-Control System): An older access-control method, common on UNIX systems, where an encrypted version of the user's log-on password must be sent across the network to an authentication server. TACACS+ is a more recent Cisco Systems Inc. proprietary approach.

Security Vulnerabilities and Exposures in Internet Systems and Services

Rui C. Cardoso
Universidade de Beira Interior, Portugal

Mário M. Freire
Universidade de Beira Interior, Portugal

INTRODUCTION

In order to guarantee a global security solution in network environments, it is necessary to take into account several issues such as security mechanisms for exchange and access to remote information; mechanisms for protection of networked systems and administrative domains; detection of new vulnerabilities and exposures; and monitoring and periodic audit of the implemented security mechanisms and disaster recovery plans.

This article is focused on the problem of detection of security vulnerabilities in an active way, using software agents. There are multiple threats to the security of computer systems and networks. The number of newly discovered vulnerabilities reported to CERT (http://www.cert.org) since 1999 continues to more than double each year. Besides, new classes of vulnerabilities are discovered each year, and subsequent reviews of existing code for examples of the new vulnerability class often lead to the discovery of evidence in hundreds of different software products. Moreover, system administrators often found themselves attacked before they even knew the existence of the vulnerability.

This article presents an overview of available software for detection of vulnerabilities and exposures in TPC/IP systems and discusses a new approach developed by the authors, based on software agents, to actively detect security vulnerabilities and exposures in Internet-based systems.

EVOLUTION OF SYSTEMS SECURITY

Network security risks are rising every day (Householder, 2002). As networks become more interconnected, the number of entry points increases and, therefore, exposes each network to threats. The widespread availability of Internet access allows the dissemination of new vulnerabilities and the know-how of hackers. While networks and applications are becoming more complex and difficult to manage, the IT industry does not appear to significantly increase the allocation of human resources to the task of securing its products. This problem is compounded by the software industry trend of shorter product lifecycles, resulting in flawed or poorly tested releases that usually have a large number of potential security weaknesses. On the other hand, hackers have suitable tools that require less technical skills and allow large-scale attacks. The time between the identification of new vulnerabilities and the exploit attempt has been reduced substantially, giving less time for administrators to patch the vulnerabilities. Moreover, hackers often have access to that information before the vendors are able to correct the vulnerabilities, in which case, it is difficult to reach the administrators in a reasonable time.

Research activities on intrusion and fault detection started in the early 1980s with the introduction of the concept of computer threats and detection of misuse by Anderson (1980). The goal of intrusion detection is simply to detect intrusions. However, intrusion detection systems (IDSs) do not detect intrusions. They only identify evidence of intrusions, either while they are in progress or after they have occurred (Manikopoulos, 2002). On the other hand, detection of security faults (holes) in hosts can anticipate the occurrence of service failures and compromises.

There are two main approaches to the problem:

1. The security companies approach mainly concentrates on the development of automated security programs capable of analyzing the attacks within a single system such as Nessus

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

(http://www.nessus.org), Nmap (http://www.nmap.org), SAINT (http://www.saintcorporation.com), SARA (http://www-arc.com/sara), and SNORT (http://snort.org). All of these software products use a standalone approach; they never share knowledge except when downloading updates from the central server. These tools used by systems administrators include databases of security vulnerabilities and exposures. However, there is a significant difference among them, and there is no easy way to determine when different databases are referring to the same problem. The consequences are potential gaps on the security coverage and no guaranty of effective interoperability among them. In addition, each tool currently uses different metrics to state the number of vulnerabilities or exposures they detect (Quo, 2002), which means there is no standardized basis for a common evaluation of these tools. The security organizations approach (Mell, 1999), followed by CVE (http://www.cve.mitre.com), ICAT (http://icat.nist.org), ISS (http://www.iss.net), NIST (http://www.nist.org), and SecurityFocus (http://www.securityfocus.org), make the publication of security alerts aimed at system administrators. In this case, it is difficult to reach the administrators in a reasonable time. Therefore, the need arises for cooperation among systems in order to manage such diverse sources of information.

2. The second approach is based on software agents for detection of vulnerabilities and exposures (Cardoso & Freire, 2003, 2004). The main objective of this approach is the development of a multi-agent system, where tasks are delegated to agents to make them cooperate with each other through agent communication language (ACL) in order to share information. This system allows the automation of tasks while minimizing the amount of needed human intervention. There are multiple threats and vulnerabilities in the security of computer systems and networks. By gathering information from those systems using software agents, it is possible to determine the nature of attacks against that networked systems.

This article briefly presents the design and implementation of an agent-based system built using JADE (Bellifemine, 1999). The main task of software agents is the detection of vulnerabilities and exposures (Cardoso, 2004; Humphries, 2000). Each agent can exchange knowledge with other agents in order to determine if certain suspicious situations actually are part of an attack. This procedure allows them to warn each other about possible threats. ICAT Metabase, a search index of vulnerabilities in computerized systems, was considered for external source of vulnerabilities used to the agent system up-to-date. The ICAT binds the users with diverse public databases of vulnerabilities as well as patch sites, thus allowing network administrators to find and repair the existing vulnerabilities in one given system. ICAT is not properly a database of vulnerabilities, but an index used by network administrators to know some reports of vulnerabilities as well as the information about patches currently available.

VULNERABILITY ASSESSMENT AND INTRUSION DETECTION

Vulnerability assessments (VA) tools automate the detection of vulnerabilities, allowing network administrators to assess the security status of their networks. These tools provide a means of detecting security holes before a malicious intruder. Some of them also provide a way to close them. Security policies, ACLs, and signed user agreements mean little, if systems are full of exploitable holes. Although host-based vulnerability assessment tools continue to be popular products, other solutions are arising. Host-based vulnerability assessment tools usually identify the version and distribution of the operating system (OS) running on a given host and test it for known vulnerabilities and exposures. Most of these tools test common applications and services on each platform. Application-layer vulnerability assessment tools are directed toward application servers. The difficulty of correctly securing a public server cannot be overstated. Most servers are exploitable due to underlying operating systems or holes in the applications. There are vulnerability assessment tools that cover more than one category. Only rarely will a host-based vulnerability assessment tool check for commonly

exploited applications on the same host on which it found the operating system. However, tools that try to do too many tasks sometimes do not perform well any specific task. A vulnerability assessment tool built specifically to analyze an application server probably will do a more thorough scan of that particular type of host than a general host scanner. All types of vulnerability assessment tools must fulfill the following roles:

1. Map the network
2. Identify the application
3. Test the vulnerabilities
4. Report the findings

The usual proceedings begin by asking either for the IP address of a specific host to scan or for the subnet range of the hosts to analyze. A well-coded vulnerability assessment tool will find all the physically connected hosts in a given network and report on the OS platform type and version (called OS fingerprinting). Vulnerability assessments will usually ping (submit an Internet Control Message Protocol—ICMP—echo) hosts, then start identifying active TCP and UDP ports. Some will automatically assume that a service running on a standard (well-known) port is a particular type of service. For instance, a weakly written vulnerability assessment tool will try to access all services running at port 80 as if it were an HTTP server, even if it is an SMTP server. Some vulnerability assessments will attempt to identify the application running on a particular port (e.g., which Instant Messaging—IM—client is active). Due to performance reasons, if the tool can recognize the application and the version, then it will only attempt the attacks that are specific for that application. There are several faults in the process of precise identification of applications and ports. Misidentifying entities could be tested for an incorrect vulnerability. There are tools that solve this problem by testing all vulnerabilities, even those that do not apply to a particular platform or port. This approach is not necessarily bad, aside from the associated performance hit. A vulnerability assessment tool should be more accurate than fast.

Some IDSs also use vulnerability assessment (sometimes referred to as scanning), which is a technology developed to assess the security of a computer system or network. They usually include:

1. Monitoring and analysis of user and system activities.
2. Analysis of system configurations and vulnerabilities.
3. Assessment of system and file integrity.
4. Ability to recognize typical patterns of attacks.
5. Analysis of abnormal activity patterns.
6. Tracking user policy violations.

Typically, an IDS system follows a two-step process. First procedures include inspection of the configuration files of the system to detect inadvisable settings; inspection of the password files to detect inadvisable passwords; and inspection of other system areas to detect policy violations. In a second step, procedures are network-based and considered an active component; mechanisms are set in place to reenact known methods of attack and to record system responses.

AVAILABLE TOOLS

There are specific software packages that are used by network administrators, such as NESSUS, Nmap, SAINT and SARA. They examine system and network configurations for vulnerabilities and exploits that unauthorized users could use. Usually they act from a specific host in the network and scan all the others. Although most of them are programs and scripts run periodically by network administrators, their use leads to a rise of processing time and to a consumption of the available bandwidth. Another negative result is that these procedures eventually may lead to an overhead on the systems performance, which may cause instability and crash in the scanned systems. Due to the lack of data exchange among those applications, there is no guarantee that all share the same knowledge.

There is a large number of tools used for VA and IDS. Table 1 shows a representative summary. A more exhaustive list can be found in the following:

1) MITRE/CVE at http://www.cve.mitre.org/compatible/product.html
2) Top 75 Security Tools at http://www.insecure.org/tools.html

Some research work has been done using Agents for Network Vulnerability Scanning (Kim, 2002; Vidal, 2003). It is based on a mobile agent approach and, therefore, if implemented correctly, can reduce the overall communication traffic in the network. Mobile agents are not always in contact with its origin host. Therefore, they have a reduced interaction when performing some tasks. They also allow the network administrator to create specialized services by tailoring the agent to a specific task. The drawback of this approach is the security issues that arise when using mobile agents in an unsecure network. This could lead to an eventual content alteration. Is the agent trustable? Is its content authenticated? An interesting approach to security in FIPA agents platforms was made by Min Zhang (Zhang, Karmouch, & Impey, 2001), which could be used to solve some of the problems previously refereed by mobile agents. Although there are few vulnerability tools using agents, there are several approaches to intrusion detection using software agents; namely, agent-based distributed intrusion detection systems (Yi, 2003) and mobile agent IDS (Humphries, 2000).

AGENT-BASED APPROACH

As more and more computers are connected to the network, the security risks increase accordingly (Manikopoulos, 2002). Although vulnerability assessment tools (Martin, 2001) (see Table 1) and IDS (Denning, 1997;; Kemmerer, 2002) are good solutions, they lack several important features (Cardoso, 2004). Network administrators need to be updated permanently and in control of everything that is appended in their networks. However, it is well known that this may not be feasible in real time. An administrator has to perform many routine and time-consuming tasks that could be delegated to software agents with advantages. The solution presented in Cardoso and Freire (2004) is based on the delegation and cooperation. By delegating tasks to specific applications capable of autonomous behavior, it is possible to enhance the overall performance of the security in a network. In that approach, agents would act as an intruder would, starting with scarce knowledge about the host and building his or her knowledge from the interaction process. This solution could provide a truer test, especially if a particular machine is exposed. Some vulnerability assessment tools require client-side agents to be installed in order to be highly efficient. Agent-based tools can be more accurate if they work on both sides of the network interface and discover processes, services, and ports that an outside service could not detect. Several vulnerability assessment tools offer both modes.

The agent-based approach uses JADE, a Java-based agent development framework (Labrou, 1999) to evaluate the feasibility of the system, as shown in Figure 1. Jade is a FIPA-Compliant Agent Platform (Bellifemine, 1999) in which Java agents can be deployed. This platform includes an AMS (Agent Management System), a DF (Directory Facilitator), and a sniffer RMA (Remote Management Agent). The available features include FIPA interaction protocols, automatic registration of agents with MAS, FIPA-compliant naming service, and FIPA-compliant IIOP (Internet Inter-Orb Protocol) to connect to different APs (Agent Platforms).

Figure 2 shows a schematic representation of the overall interconnection solution based on distributed SAPA Agents (Software Agents for Prevention and Auditing of Security Faults in Networked Systems). A SAPA Agent receives solicitations to perform a specific task. The tasks currently supported are the following: host scan; network scan; and host/network monitoring, in which the agent performs a detailed scan of the network, the open ports, and the services active in those ports. Thereafter, it collects all the requested data. It will use data from previous scans and data stored from several external sources to build a knowledge base. These sources are the ICAT database of known vulnerabilities and exposures, the PortsDB (http://www.portsdb.org) that tells us which services are associated with specific ports. After the inference process is complete and the agent has gathered sufficient knowledge about the situation, it

Figure 1. SAPA agent in a JADE platform

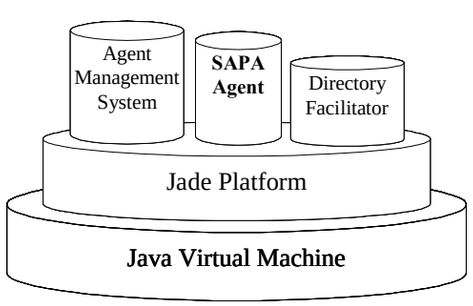
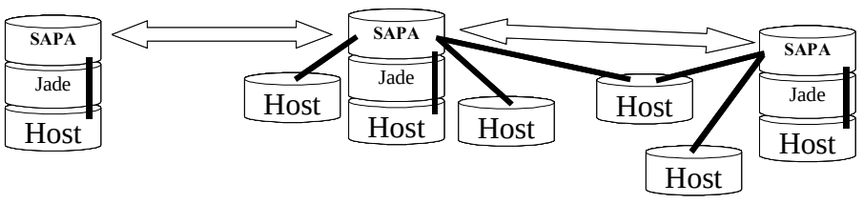


Table 1. Vulnerability assessment appliances

Entity	Appliance	Type
CERT Vulnerability Database	CERT Coordination Center	Archives / Database
Cert-IST knowledge base	Cert-IST	Vulnerability Database
Check Point VPN-1/FireWall-1 with SmartDefense	Check Point Software Technologies, Ltd.	Scalable Small Office to Enterprise VPN and Firewall
Cisco Secure / Secure IDS	Cisco Systems	Vulnerability Database / Intrusion Detection System
Dragon Sensor	Enterasys Networks	Packet-Based Intrusion Detection System
DragonSoft Secure Scanner	DragonSoft Security Associates, Inc.	Vulnerabilities and Exposures Assessment Software
Foundstone Enterprise Risk Solutions	Foundstone, Inc.	Managed Security Assessment Service
ICAT	National Institute of Standards and Technology	Vulnerability Database
iNETPATROL	Network Security Systems	Vulnerability Testing and Reporting Service
McAfee Entercept / IntruShield IDS	McAfee, Inc.	Host Protection System / Network-Based Intrusion Detection System
Nessus Security Scanner	The Nessus Project (Renaud Deraison & Jordan Hrycaj)	Vulnerability Assessment Tool
Netcraft Network Examination Service	Netcraft Ltd.	Managed Vulnerability Scanning Service
netForensics	netForensics, Inc.	Security Information Management
Open Source Vulnerability Database	Open Source Vulnerability Database (OSVDB)	Vulnerability Database
OVAL (Open Vulnerability Assessment Language) Web site	MITRE Corporation	Standard for Describing Vulnerability Presence Criteria Web site
PatchAdvisor Enterprise	PatchAdvisor, Inc.	Patch Management
PatchAgent	NISCENT s.r.l.	Patch Management Tool
Penetration Study	Clear North Technologies, Inc.	Penetration Study
QualysGuard	Qualys	Network and Application Vulnerability Assessment Platform
QualysGuard SANS/FBI Top 20 Vulnerabilities Scanner	Qualys	Free Vulnerability Assessment Service
Retina Network Security Scanner	eEye Digital Security	Vulnerability Assessment Tool
SAINT / SAINTbox	SAINT Corporation	Vulnerability Assessment Tool / Network Vulnerability Scanning Appliance
SANS GIAC Security Training	SANS Institute	Educational Material
SecurityFocus Vulnerability Database	Symantec	Vulnerability Database
SecuritySpace Security Audits	E-Soft, Inc.	Vulnerability Assessment Service
SecurityTracker	SecurityTracker	Vulnerability Alerts
Snort	Snort Development Team	Intrusion Detection System
Symantec Security Response Web site	Symantec	Vulnerability Database, Security Advisories and Archives
Symantec Vulnerability Assessment	Symantec	Network Vulnerability Assessment
System Scanner 4.2	Internet Security Systems, Inc. (ISS)	Host Vulnerability Assessment Tool
Trend Micro Vulnerability Assessment	Trend Micro, Inc.	Vulnerability Assessment Product With Virus Info Association
WebSAINT	SAINT Corporation	Web-Based Vulnerability Scanning Service
X-Force Alerts and Advisories /Database	Internet Security Systems, Inc. (ISS)	Advisory Archive/Database

Figure 2. SAPA MAS multi-agent system



will create an output result. Finally, the process is finished when the output is presented to the requester. Figure 2 presents a brief overview of the multi-agent system based on SAPA agent interactions.

CONCLUSION

This article presented an overview of detection of security vulnerabilities and exposures in networked systems. The problem of vulnerability assessment has been addressed. An analysis of the currently available software tools in VA/IDS was presented, and an overview of a new approach based on software agents was briefly discussed.

REFERENCES

Anderson, J.P. (1980). *Computer security threat monitoring and surveillance.* Fort Washington, PA: James P. Anderson, Co.

Bellifemine, F., et al. (1999). JADE—A FIPA-compliant agent framework. *Proceedings of PAAM99*, London.

Cardoso, R.C., & Freire, M.M (2003). An agent-based approach for detection of security vulnerabilities in networked systems. *Proceedings of 11th International Conference on Software, Telecommunications and Computer Networks (SoftCom'2003)*, Dubrovnik, Croatia) and Venice, Italy.

Cardoso, R.C., & Freire, M.M. (2004a). Intelligent assessment of distributed security in TCP/IP networks. In Z. Mammeri, & P. Lorenz (Eds.), *High-speed networks and multimedia communications* (pp. 1092-1099). Berlin Heidelberg: Springer-Verlag.

Cardoso, R.C., & Freire, M.M (2004b). FIPA-compliant software agents for detection of vulnerabilities and exposures in TCP/IP hosts. *Proceedings of the International Conference on Information Networking ICOIN 2004*, Busan, Korea.

Denning, D.E. (1987, February). *An intrusion-detection model. IEEE Transactions on Software Engineering*.

Householder, A., Houle, K., & Dougherty, C. (2002, April). Computer attack trends challenge Internet security. *IEEE Computer, Security and Privacy—Supplement*, 5-7.

Humphries, J.W., & Pooch, U.W. (2000). Secure mobile agents for network vulnerability scanning. *Proceedings of the 2000 IEEE Workshop on Information Assurance and Security*, New York.

Kemmerer, R.A., & Vigna, G. (2002, April). Intrusion detection: A brief history and overview. *IEEE Computer, Security and Privacy—Supplement*, 27-29.

Kim, B., Jang, J., & Chung, T.M. (2002). Design of network security control systems for cooperative intrusion detection. In I. Chong (Ed.), *Information networking* (pp. 389-398). Heidelberg: Springer Verlag.

Labrou, Y., Finin, T., & Peng, Y. (1999). Agent communication languages: The current landscape. *IEEE Intelligent Systems, 14*(2), 45-52.

Manikopoulos, C., & Papavassiliou, S. (2002). Network intrusion and fault detection: A statistical anomaly approach. *IEEE Communications Magazine, 40*(10), 76-82.

Martin, R.A. (2001). Managing vulnerabilities in networked systems. *IEEE Computer, 34*(11), 32-38.

Mell, P. (1999). Understanding the world of your enemy with I-CAT (Internet-categorization of attacks toolkit). *Proceedings of the 22nd National Information System Security Conference*, Arlington, Virginia, USA.

Quo, G., Rudraraju, J., Modukuri, R., & Hariri, S. (2002). A framework for network vulnerability analysis. *Procedings of the IASTED International Conference Communication, Internet & Information Technology*, St. Thomas, Virgin Islands, USA.

Vidal, J.M., & Pedireddy, T. (2003). A prototype multiagent network security system. *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS'03*, Melbourne, Australia.

Yi, M.-K., & Hwang, C.-S. (2003). Design of fault tolerant architecture for intrusion detection systems using autonomous agents. *Proceedings of the International Conference on Information Networking (ICOIN'2003)*, Jeju, Korea.

Zhang, M., Karmouch, A., & Impey, R. (2001). Towards a secure agent platform based on FIPA. *Proceedings of IEEE/ACM MATA 2001,* Montreal, Canada, 277-289.

KEY TERMS

Agent: A component of software and/or hardware that is capable of acting in order to accomplish tasks on behalf of its user. Software agents are agents in the form of programs (code) that operate in computer environments.

CVE: Common Vulnerabilities and Exposures is a list of standardized names for vulnerabilities and other information security exposures. CVE aims to standardize the names for all publicly known vulnerabilities and security exposures.

FIPA: Foundation for Intelligent Physical Agents, an international non-profit association of companies promoting and developing specifications to support interoperability among agents and agent-based applications (http://www.fipa.org).

ICAT: Internet Categorization of Attacks Toolkit.

IDS: Intrusion Detection System (IDS) is a type of security management system for computers and networks in which the IDS gathers and analyzes information from various areas within a computer or a network in order to identify possible security breaches, which include both intrusions (attacks from outside the organization) and misuse (attacks from within the organization).

JADE: Java Agent DEvelopment framework, a middleware tool used in implementation of agent-based systems (http://jade.tilab.com).

MAS: Multi-Agent Systems are systems in which many intelligent agents interact with each other. The agents are considered to be autonomous entities, such as software programs or robots. Their interactions can be either cooperative or selfish.

Software Agents: Sometimes also known as bots, protect their users from the complexity of computer and network operations, and may engage in database searches and transactions based upon knowledge of an evolving user profile. For more information, see the UMBA Agents' Web site http://agents.umbc.edu/ .

VA: Vulnerability Assessment is an examination process of the ability of a system or application, including current security procedures and controls, to withstand an intrusion. A vulnerability assessment may be used to identify weaknesses that could be exploited and predict the effectiveness of additional security measures in protecting information resources from attack.

Semantic Web

Rui G. Pereira
Universidade de Beira Interior, Portugal

Mário M. Freire
Universidade de Beira Interior, Portugal

THE WEB

The World Wide Web (WWW, Web, or W3) is known as the largest accessible repository of human knowledge. It contains around 3 billion documents, which may be accessed by more than 500 million worldwide users. In only 13 years since its appearance in 1991, the Web suffered such a huge growth that it is safe to say there is no phenomenon in history that can compare to it. It reached such importance that it became an indispensable partner in the lives of people (Daconta, Obrst & Smith, 2003).

Researching information on the current Web is supported through the use of robust and practical applications known as search engines and directories. But the fast and unorganized growth of the Web is making it difficult to locate, access, present, and maintain online trustful content for an increasing number of users. Difficulties in the search of Web contents are associated with the use of non-structured, sometimes heterogeneous information, and with the ambiguity of Web content. Thus, one of the limitations of the current Web is the lack of structure of its documents and the information contained in them. Besides, information overload and poor aggregation of contents make the current Web inadequate for automatic transfers of information (Berners-Lee, Hendler & Lassila, 2001; Lu, Dong & Fotouhi, 2002; Moura, 2001).

Another limitation is the fact that the current Web uses only a human-oriented type of communication. Information on the Web is conceived to have human beings as its consumers, not to be understandable by machines or software agents. In these circumstances, only human beings can understand and manipulate online information. Because of this, the current Web is not exploring all of its potentials (Berners-Lee, Hendler & Lassila, 2001).

Berners-Lee et al. (2001) appear as founders of the next generation Web, to which is given the name Semantic Web. The idea of the Semantic Web is to bring the Web to its full potential. It will have a positive impact on all levels, from individual users to large companies.

THE SEMANTIC WEB

The vision of the Semantic Web, as proposed by Berners-Lee et al. (2001), is the evolution of the current Web to one where data and services are understandable and usable by humans as well as computers. It is important to say that the Semantic Web is not artificial intelligence (Berners-Lee, 1998). The objective of the Semantic Web is not to make computers understand the human language. Processing and relating of contents do not mean an intelligent processing in the same concept that is used by the artificial intelligence researchers. The challenge of researchers of the Semantic Web is to define a universal language for the expression of data and a set of inference rules that computational agents can use to process it.

The idea of the Semantic Web is to define and link the data in a way that it can be used by machines, not just for display purposes, but for automation, integration, and reuse across heterogeneous applications. For that purpose, data needs to be application-independent, classifiable, editable, and part of a large information ecosystem (known as ontology). This type of data, embodied with semantic information, is known as smart data. With smart data, machines can interpret, manipulate, and make inferences about it. The Semantic Web can be defined as a machine-processable Web composed of smart data; the power is moving from applications to data (Berners-Lee,

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Table 1. Main advantages of the Semantic Web

Main Advantages of the Semantic Web
• It is a way to describe and expand the current Web, adding a concept layer to it. • Allows machine-readable, interpretable and editable web content. • Offers a way to enable semantic annotations that could be easily organized and found. • Enhances search mechanisms with the use of "Ontologies" - relationships and axioms between concepts, allowing the standardization of web annotations, service descriptions, and web data to be meaningfully related. • Enables software agents to carry sophisticated tasks automatically – through the use of smart data. • Allows better communication between platform-independent software agents. • Enables the use of levels of trust for information.

Hendler & Lassila, 2001; Daconta, Obrst & Smith, 2003; Zhao & Sandahl, 2003). The main advantages of the Semantic Web can be grouped into several points, as shown in Table 1.

As the current Web allows the sharing of documents between previously incompatible computers, the Semantic Web intends to go beyond, allowing stovepipe systems, hardwired computers, and other devices to share document embedded contents (Berners-Lee, Hendler & Lassila, 2001; Daconta, Obrst & Smith, 2003).

Although very promising, the implementation of the Semantic Web is of enormous complexity. The first challenge consists of establishing standards that define an intelligent and universal form of content of Web pages in order to support better interpretation by the machines. The second stage consists of developing programs that obtain and share data from several sources. Past these two stages, it will become necessary to develop software agents that can generate new information based on the available one (Moura, 2001).

The Semantic Web is on its first steps. Therefore, its ideology and current structure must be seen as the base for future development of the Web and not as a final prototype. Researchers should avoid predictions that are too optimistic or incorrect, as happened with the first researchers of artificial intelligence in the 1950s and 1960s. It is important that Semantic Web researchers keep their feet on the ground.

Other point of attention is the resistance that human beings have to changes, new ideas, and, in particular, to having a complete understanding of the Semantic Web. The Semantic Web needs the network effect to become a reality. It is very important to alert people that a single Semantic Web page does not have any power by itself. In fact, the development of integrated tools for the Semantic Web that could be easily used and tested by every Web user is still not a reality. Research activities for these new tools need to take into account that the way people are currently writing Web pages must be changed as little as possible.

Another point is the implementation of the Semantic Web, which can create more problems than it is trying to solve. To be successful, the Semantic Web needs to be simple and powerful for all Web users. According to Clark (2002), "the Semantic Web is most often criticized by its detractors for having, in the end, very little to do with reality; or, put less pointedly, for being easier to dream about than to implement" (p. 1). Champion (2002) suggested three general threads in the critiques of the Semantic Web initiative: first, W3C members have a higher priority of Web services interoperability than Semantic Web affords; second, to date, progress of Semantic Web efforts has not shown any really useful example for Web users; third, the Semantic Web vision, as other visions in recent history, is unlikely to live up to its promise. Even though the Semantic Web may yet seem a remote dream, researchers believe that the Semantic Web will be adopted by users in the future—real conclusions on the viability of the Semantic Web are still in years to come.

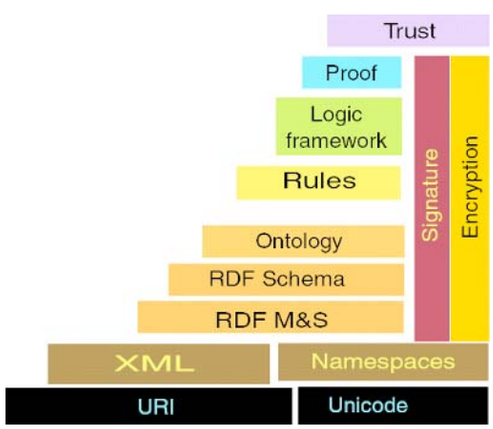
THE STRUCTURE OF THE SEMANTIC WEB

The structure of the Semantic Web presented by W3C is based on a Semantic Web stack, as shown in Figure 1. As we move up in the stack, technologies represent

information in a more expressive and meaningful way. Not all technologies of this stack are standardized yet, but the prototypes are promising. The stack can be split into three parts: Fully Standardized (bottom layers: URI/Unicode and XML/Namespaces); Standardization in Progress (medium layers: RDF M&S, RDF Schema and Ontologies); and Still Experimental (upper layers: Rules, Logic Framework, Proof, and Trust).

In the bottom layers, the base layer of the stack is URI, the concept of Uniform Resource Identification, and Unicode, a universal character set. The Semantic Web uses the concept of resource. All the documents in the current Web, as well as objects, concepts, or events of the real world, are defined as resources. People, books, places, or ideas are some examples of resources in the Semantic Web. The URI is a string that unambiguously identifies a resource. Unicode is a standard for representing characters as integers. Unicode contains 34,168 distinct coded characters derived from 24 supported language scripts. These characters cover the principal written languages of the world. The use of URIs and Unicode together preserves the use of international character sets and enables the Semantic Web resource identification (Koivunen & Miller, 2001).

Figure 1. The semantic Web stack



(http://www.w3.org/DesignIssues/diagrams/sw-stack-2002.png)

Copyright © 2002 World Wide Web Consortium, (Massachusetts Institute of Technology, European Research Consortium for Informatics and Mathematics, Keio University). All Rights Reserved. http://www.w3.org/Consortium/Legal/2002/copyright-documents-20021231

Above URI and Unicode, we find the XML and Namespaces layer. XML, one of the support stones in the Semantic Web, is a set of rules for structuring Web documents and data. Increasingly, XML plays a role of mechanism of exchange and interoperability between different applications. Because it is human-readable plaintext, XML is independent from applications and operation systems. It is easily readable and understandable by a variety of software agents and systems that will need to consume data on the Web (Ahmed et al. 2001).

XML is becoming the glue that binds all Semantic Web technologies together, the base of semantic representation and processing (Daconta, Obrst & Smith, 2003). Table 2 shows the main advantages of XML.

Namespaces is a simple mechanism for creating globally unique names in domain-specific vocabularies. It is used to distinguish identical names in different markup languages, allowing domain-specific names to be mixed together without ambiguity. Above the XML and Namespaces layer are the medium layers of the Semantic Web stack. The technologies in these layers—RDF, RDF Schema, and Ontologies—are still in standardization process.

RDF is the concept of Resource Description Framework, a W3C recommendation since 1999. It is an XML-based language that uses a triple-based assertion model and a syntax to describe resources. RDF model is called "triple," because it can be described in terms of subject, predicate, and object, like grammatical parts of a sentence (Beckett, 2001; Hayes 2002; Manola & Miller; 2002), as shown in Figure 2.

The basic RDF data model consists of four object types:

- **Resources:** All that can be described by an RDF expression;
- **Properties:** Specific aspects, characteristics, attributes, or relations used to describe a resource;
- **Literals:** Constant values represented as character strings; and
- **Statements:** Combines resources, properties and property values together (RDF triple).

Table 2. Main advantages of XML

Main Advantages of XML
• Simplifies the creation of application-independent documents and data. • Allows organization of document contents in a single hierarchical tree that can be easily validated. • Acts as a standard structure to separate information data from presentation data. • Provides a simple standard syntax for providing information about data values.

Figure 2. A graph of one statement (Triple model)

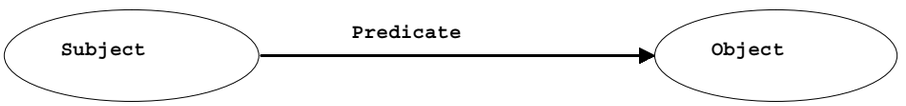
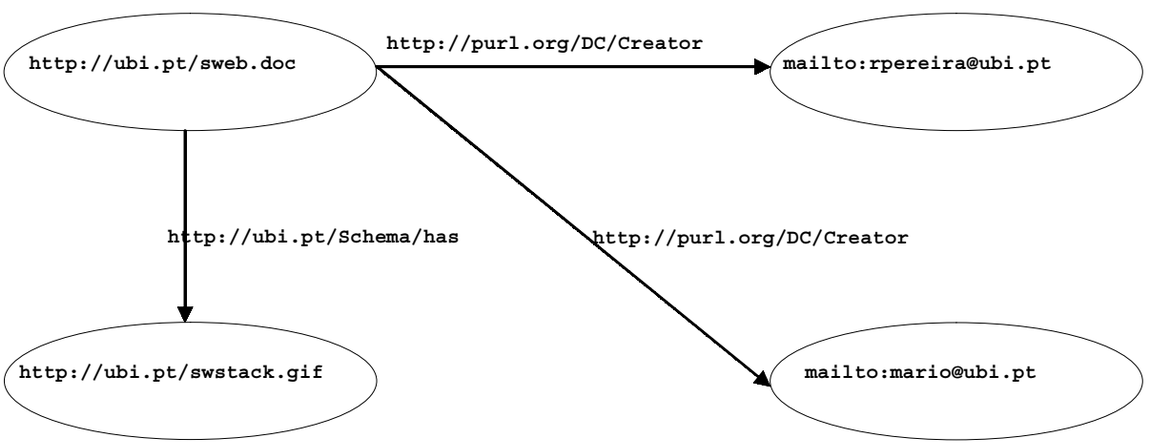


Figure 3. A graph of three RDF statements (RDF Triples)



An RDF document contains one or more RDF descriptions—sets of RDF statements about a resource, as exemplified in Figure 3. Most RDF authors write their RDF assertions in Notation 3 format and then convert them to RDF/XML syntax via a conversion tool (i.e., HP Jena) application. Notation 3 is an equivalent language to RDF/XML, but it is simpler, more compact, and readable (Daconta, Obrst & Smith, 2003).

Above the RDF layer, we have RDF Schema. RDF Schema expresses a hierarchy-class data model used for the classification and description of standard RDF resources. The role of RDF Schema is to facilitate the definition of metadata by providing a data model, much like many object-oriented programming languages, in order to allow the creation of data classes. It is a simple language that enables people to create their own RDF vocabularies in RDF/XML syntax (Brickley & Guha, 2002).

Above RDF Schema layer, we have the Ontology layer. Ontology is a general description of all concepts or meanings (what we call "semantics") and their relationships inside a particular area of knowledge, like medicine, mathematics, automobile repair, and so forth. Any ontology can be easily extended, refined, and reused by other ontologies, providing expressive representation for a wide diversity of concepts from the real world. Ontologies can go through several degrees of semantic richness. In its simplest form, an ontology can represent knowledge with a minimal hierarchic structure, like a taxonomy. However, in more complex forms, ontology can represent very

rich, complex, consistent, and meaningful knowledge, like a thesaurus (words, synonyms and their relationships), conceptual model, and logical theory (Berners-Lee et al., 2003).

A conceptual model is a model of a specific area of knowledge that represents entities (concepts or events) and relations among them. Each entity has a set of attributes. Simple rules like axioms (set of statements asserted to be true) or constraints also can be applied.

Logical theories are supported by axioms and inference rules. Inference rules allow an assumption-based generation of valid conclusions. Axioms and inference rules are used to prove theorems surrounding the represented area of knowledge. The whole set of axioms, inference rules, and theorems constitute the logical theory. Ontologies represented as logical theories have high semantic richness and are semantically interpretable by machines. Logical theories are the state of the art in the knowledge representation of the Semantic Web (Daconta & Obrst & Smith, 2003).

Above the ontology layer are the upper layers of the Semantic Web Stack. Technologies and concepts in these layers—Rules, Logic Framework, Proof, and Trust—are still in academic discussion and research. In the Rules layer, we can define logic rules about concepts represented in the ontologies. Rule-based languages are a powerful way of expressing relationships. Because ontologies represent concepts and their relationships in a logical and machine-interpretable form, automated rules can infer new knowledge from the primary knowledge of the ontology. Additionally, technology in this layer can provide a standard way to query and filter RDF documents. While the Rules layer uses simple logic capability, the Logic Framework layer uses advanced logic capability to embed axioms on rule-based systems. Once systems are built to follow logic, they can be used to prove things. Everyone could write logic statements that machines could follow to construct proofs (new valid knowledge). Logic Framework also allows formal logic proofs to be shared (Swartz, 2002).

With robust proofs, a Trust layer can be established. This "web of trust" form is the W3C three-part vision of the Web (a collaborative Web, a Semantic Web, a web of trust). In a web of trust, machines can evaluate the trustfulness of RDF statements. Like in the current Web, not all the information in the Semantic Web will be trustworthy.

Obtaining levels of trust for information is one of the objectives of the Semantic Web. Trustfulness will be evaluated locally by each Semantic Web application. Documents classified as trustful will be preferred by Semantic Web applications. Levels of trust accepted by Semantic Web applications can be configured (Berners-Lee, Hendler & Lassila, 2001; Swartz, 2002).

XML signature and XML encryption are applicable to technologies above the XML and Namespace layer. Information can be digitally signed and encrypted using the XML-signature specification and XML-encryption to assure the authenticity and integrity of information (Berners-Lee, Hendler & Lassila, 2001; Swartz, 2002).

Semantic Web agents are applications that can use all Semantic Web layers to display more intelligent behavior, capable of offering more intelligent services, including possibilities that may not have been considered yet. Because Semantic Web pages are meaningful for machines, Semantic Web agents can be simpler and more powerful than current ones (Berners-Lee, Hendler & Lassila, 2001).

Most of the Semantic Web tools available are used for research. Table 3 shows some of the most successful tools for the Semantic Web.

CONCLUSION

Despite being based on previous information retrieval and knowledge representation projects, the Semantic Web goes beyond them. The Semantic Web shows a way to define, add, extract, and contextualize real-world concepts without ambiguity for various areas of knowledge in a non-centralized way, the beginning of a machine interpretable Web, usable by users and applications (Berners-Lee, Hendler & Lassila, 2001; Gandon, 2002).

The Semantic Web values information, enabling applications to extract knowledge from data in a simple way. The Semantic Web is a pioneer in acknowledge that information is more important than applications. Data revolution starts with information being highly structured and classified, application-independent, and optimized for machine interpretability. Introduction of self-descriptive data contextualized in knowledge areas (ontologies) means

simpler, smarter applications that are more able to be understood and can handle complex tasks.

The Semantic Web consequence is that machines will interpret information the way we do, bringing machines to our communication level. Global development and adoption of ontologies are steps toward smart machines being able to process information at our conceptual level. A machine's ability to interpret reality is clearly still very primitive, but on the Semantic Web, they can easily interact, learn, and evolve.

The Semantic Web is no fiction, but it is not reality, either—it is a real possibility and a big challenge, as the Web itself was in its beginnings. The Semantic Web will integrate, interact, and bring benefits to all human activities. Its full potential is to go beyond the Web to real-world machines, providing increased machine-to-machine and machine-to-human interaction. Phones, radios, and other electronic devices will intercommunicate using the Semantic Web as standard. The Semantic Web is one more step in a human-like form of approach of the machines to reality and in the evolution of human knowledge.

REFERENCES

Ahmed, K., et al. (2001). *XML meta data*. Birmingham, UK: Wrox Press, Ltd.

Alexaki, S., et al. (2002). The ICS-FORTH RDFSuite: High-level scalable tools for the Semantic Web. *ERCIM News, 51*. Retrieved August 2004, from *http://www.ercim.org/publication/Ercim_News/enw51/alexaki.html*

Table 3. Some Semantic Web tools

Some Semantic Web Tools
• **Jena Java RDF API and toolkit:** A Java RDF framework developed by Hewlett-Packard researchers for writing Semantic Web applications. It contains an API for manipulating RDF models, including statement and resource-centric methods using cascading calls for easy object orientated use, container support, the ARP RDF/XML parser, an RDF/XML writer, the RDQL query language, DAML support, and persistent storage (Wilkinson & Sayers & Kuno & Reynolds, 2003). • **ICS-FORTH RDFSuite:** High-level scalable tools for the Semantic Web. (Alexaki et al., 2002). • **IsaViz:** A visual environment for browsing and authoring RDF models represented as graphs (http://w3c.org/2001/11/IsaViz/). • **Metalog:** A next-generation reasoning system for the Semantic Web, which has been designed to be particularly user-friendly and easy-to-use in order to showcase the initial potential of the Semantic Web. It is entirely written in Python and can be easily interfaced with every logic programming system (http://www.w3.org/RDF/Metalog/). • **OntoEdit.** A modeling and administration framework for ontologies and ontology-based solutions for integration of heterogeneous structures, patterns, and models (http://www.ontoprise.de/home_en). • **OWL – Web Ontology Language:** The most expressive of ontology languages currently defined for the Semantic Web. It has been developed by the W3C's Web Ontology Working Group and intended to be the successor of the DAML+OIL language (Dean et al., 2002). • **Protégé:** An ontology management tool developed and maintained by the Medical Informatics Laboratory at Stanford University. It is recognized as an exemplary tool for managing ontologies (Noy et al., 2000). • **RDFDB:** A database that directly supports the RDF data model. It can load data from files, or data can be inserted using the database API. It also supports an SQL-like query language (http://rdfdb.sourceforge.net/). • **Sesame:** A Java-based storage and querying middleware system for RDF and RDF Schema. It contains an API for RDFs' manipulations on repositories, an RDF Model Theory inference engine, and support for the RQL and RDQL query languages (Broekstra et al., 2002).

Beckett, D. (2004). RDF/XML syntax specification (revised) [W3C working draft]. Retrieved August 2004, from *http://www.w3.org/TR/rdf-syntax-grammar/*

Berners-Lee, T. (1998). What the Semantic Web can represent. *W3C*. Retrieved August 2004, from *http://www.w3.org/DesignIssues/RDFnot.html*

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *The Scientific American*. Retrieved from *http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21*

Brickley, D., & Guha, R.V. (2002). RDF vocabulary description language 1.0: RDF schema [W3C working draft]. Retrieved from *http://www.w3.org/TR/rdf-schema/*

Broekstra, J., Kampman, A., & van Harmelen, F. (2002) Sesame: A generic architecture for storing and querying RDF and RDF schema. *Proceedings of the First International Semantic Web Conference*. Sardinia, Italy.

Champion, M. (2002). SemWeb again posts. *ActiveState Programmer Network*. Retrieved from *http://aspn.activestate.com/ASPN/Mail/Message/1182403*

Clark, G.K. (2002). If ontology, then knowledge: Catching up with WebOnt. *O'Reilly XML.com*. Retrieved from *http://www.xml.com/pub/a/2002/05/01/webont.html*

Daconta, M.C., Obrst, L.J., & Smith, K.T. (2003). *The Semantic Web*. Indianapolis, Indiana: Wiley Publishing, Inc.

Dean, M., et al. (2002). OWL Web ontology language 1.0 reference [W3C working draft]. Retrieved August, 2004 from *http://www.w3.org/TR/2002/WD-owl-ref-20021112/*

Gandon, F. (2002). Distributed artificial intelligence and knowledge management: Ontologies and multi-agent systems for a corporate Semantic Web [scientific philosopher doctorate thesis]. INRIA and University of Nice, Sophia Antipolis: Doctoral School of Sciences and Technologies of Information and Communication (S.T.I.C.).

Hayes, P. (2002). RDF semantics [W3C working draft]. Retrieved from *http://www.w3.org/TR/rdf-mt/*

Koivunen, M., & Miller, E. (2001). W3C Semantic Web activity. *Proceedings of the* Semantic Web Kick-off Seminar, *Finland*.

Lu, S., Dong, M., & Fotouhi, F. (2002). The Semantic Web: Opportunities and challenges for next-generation Web applications. *Information Research, 7*(4). Retrieved August 2004, from *http://InformationR.net/ir/7-4/paper134.html*

Manola, F., & Miller, E. (2002). RDF primer [W3C working draft]. Retrieved August 2004, from *http://www.w3.org/TR/2002/WD-rdf-primer-20020319/*

Moura, A. (2001). A Web semântica: Fundamentos e tecnologias. *Proceedings of the CICC 2001—Congreso Internacional de Ciencias de la Computación*. Bolívia.

Noy, N.F., Fergerson, R.W., & Musen, M.A. (2000). The knowledge model of Protege-2000: Combining interoperability and flexibility. *Proceedings of the 2nd International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, Juan-les-Pines, France.

Swartz, A. (2002). The Semantic Web in breadth. Retrieved from *http://logicerror.com/semanticWeb-long*

Wilkinson, K., Sayers, C., Kuno, H., & Reynolds, D. (2003). Efficient RDF storage and retrieval in Jena2. *Proceedings of the First International Workshop on Semantic Web and Databases*, Berlin, Germany.

Zhao, Y., & Sandahl, K. (2003). Potential advantages of Semantic Web for Internet commerce. *Proceedings of the International Conference on Enterprise Information Systems (*ICEIS'2003*)*, Angers, France.

KEY TERMS

Metadata: In some English words, the prefix *meta* indicates "change"; in others, including those related

to data and information, the prefix carries the meaning of "underlying definition or description." XML is sometimes referred to as metadata, because it describes how to represent a collection of data. Therefore, metadata means description of data. On the Web, metadata is used to describe Web resources and how to structure them.

Namespace: A simple mechanism for creating globally-unique names in domain-specific vocabularies. It is used to distinguish identical names in different markup languages, allowing domain-specific names to be mixed together without ambiguity. Each namespace is identified by a URI reference and easily can be used in XML documents.

Ontology: The word *ontology* comes from the Greek *ontos* (being) and *logia* (written or spoken discourse). It has been in use since Empedocles described the four elements—air, earth, fire, and water. It was reintroduced by 19th-century German philosophers to distinguish the study of various kinds of beings in the natural sciences. The word *ontology* can be and has been used with very different meanings. In artificial intelligence, ontology is defined as a working model of concepts and interactions from a particular domain of knowledge (e.g., medicine, mathematics, automobile repair, etc.), which is used to easily describe the meaning of different contents that can be exchanged in information systems. Any ontology easily can be extended, refined, and reused by other ontologies, providing expressive representation for a wide diversity of real-world concepts. Semantic-rich ontologies are the state-of-the-art in the knowledge representation of the Semantic Web.

Stovepipe Systems: A system where all the components are hardwired only to work together.

URI – Uniform Resource Identification: Also known as Universal Resource Identifier, it is a string that unambiguously identifies a resource. A URI describes the mechanism used to access the resource, the specific computer where the resource is housed, and the specific name of the resource. URL (Uniform Resource Locater), the most common form of URI, and URN (Uniform Resource Name) are subsets of the URI.

W3C – World Wide Web Consortium: An industry consortium that was created in October 1994 to lead the Web to its full potential by developing common protocols that promote its evolution and ensure its interoperability. The consortium is international, jointly hosted by the MIT Laboratory for Computer Science in the United States, ERCIM (European Research Consortium in Informatics and Mathematics), and Keio University of Japan, which are responsible for core development and local support. The W3C has around 350 member organizations from all over the world and has earned international recognition for its contributions to the growth of the Web. W3C was established initially in collaboration with CERN (European Organization for Nuclear Research), DARPA (Defense Advanced Research Projects Agency), and the European Commission.

XML – Extensible Markup Language: A standard with very flexible and simple syntax (a small set of rules in human-readable plaintext) used to describe and share commonly structured platform-independent information. Main components of its structure are elements (markups) and attributes of elements that are nested to create a hierarchical tree that easily can be validated. XML is extensible, because, unlike HTML, anyone can define new tags and attribute names to parameterize or semantically qualify contents. It has been a formal recommendation from W3C since 1998, playing an increasingly important role in the exchange of a wide variety of data on the Web.

Software Ad Hoc for E-Learning

Maria-Isabel Sánchez-Segura
Carlos III Technical University of Madrid, Spain

Antonio de Amescua
Carlos III Technical University of Madrid, Spain

Luis García
Carlos III Technical University of Madrid, Spain

Luis A. Esteban
Carlos III Technical University of Madrid, Spain

THE SOURCE OF THE PROBLEM

Although they are non-educational institutions, financial institutions have specific training needs. The greatest priority in employee training arises when the bank launches a new financial product or service. The difficulty, in such cases, lies in training the employees in all the regional branches so that they can offer good service to meet the clients' demand for the product.

In developing the training program two factors had to be considered:

- The department responsible for developing the new financial product keeps it secret during the development phase. Therefore, the technical details, tax treatment, and other issues relating to the product are known only after it has been designed and is ready to be launched. Consequently, it is impossible to train employees until the new product has been completely developed.
- Traditionally, employee training is pyramidal. First of all, the trainers in each training center are trained. These, in turn, train the managers, in groups, from the most important branches. Finally, these managers are responsible for training the employees in their offices.

Considering the specific needs of the employees, and to obtain the maximum profitability from new financial products, we defined the pilot project called factory to minimize time and cost spent in the development of e-learning courses for financial institutions.

This project was conceived to cover the above-mentioned weaknesses detected in the training process of an important financial institution. The pilot project goals were:

- To improve the spread of knowledge
- To minimize the course development cost and time

This pilot project consisted of two main parts: developing the factory tool and developing the courses with and without this tool, in order to compare the cost/benefit for the institution.

E-LEARNING

E-learning, also known as "Web-based learning" and "Internet-based learning", means different things to different people. The following are a few definitions of e-learning:

- E-learning is the convergence of learning and the Internet. (Bank of America Securities)
- E-learning is the use of network technology to design, deliver, select, administer, and extend learning. (Elliott Masie, The Masie Center)
- E-learning is Internet-enabled learning. Components can include content delivery in multiple formats, management of the learning experience, and a networked community of learners, content developers and experts. E-learning provides faster learning at reduced costs, increased

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

access to learning, and clear accountability for all participants in the learning process. In today's fast-paced culture, organizations that implement e-learning provide their work force with the ability to turn change into an advantage. (Cisco Systems)

- E-learning is the experience of gaining knowledge and skills through the electronic delivery of education, training, or professional development. It encompasses distance learning and asynchronous learning, and may be delivered in an on-demand environment, or in a format customized for the individual learner (Stark, Schmidt, Shafer, & Crawford, 2002).
- E-learning is education via the Internet, network, or standalone computer. Network-enabled transfer of skills and knowledge. E-learning refers to using electronic applications and processes to learn. E-learning applications and processes include Web-based learning, computer-based learning, virtual classrooms, and digital collaboration. Content is delivered via the Internet, intranet/extranet, audio or video tape, satellite TV, and CD-ROM (E-learnframe, 2004).
- Any technologically mediated learning using computers whether from a distance or in face to face classroom setting (computer assisted learning) (USD, 2004).
- Any learning that utilizes a network (LAN, WAN or Internet) for delivery, interaction, or facilitation. This would include distributed learning, distance learning, computer-based training (CBT) delivered over a network, and WBT. It can be synchronous, asynchronous, instructor-led or computer-based, or a combination (LCT, 2004).

In a general way, the most accepted definition for e-learning is: "The use of technologies to create, distribute and deliver valuable data, information, learning, and knowledge to improve on-the-job and organisational performance and individual development." Although it seems to focus on Web-based delivery methods, it is actually used in a broader context.

There are many well-known organizations that are making a big effort to standardize the concepts, processes and tools that have been developed around e-learning:

- The Aviation Industry CBT (Computer-Based Training) Committee (AICC) (http://www.aicc.org/) (AICC, 1995, 1997), is an international association of technology-based training professionals. The AICC develops guidelines for the aviation industry to develop, deliver, and evaluate CBT and related training technologies. The AICC develops technical guidelines, (known as AGR's), for example, Platform guidelines for CBT delivery (AGR-002), a DOS-based digital audio guideline (AGR-003) before the advent of window multimedia standards, a guideline for Computer Managed Instruction (CMI) interoperability, this guideline (AGR-006) resulted in the CMI systems that are able to share data with LAN-based CBT courseware from multiple vendors. In January 1998, the CMI specifications were updated to include Web-based CBT (or WBT). This new Web-based guideline is called AGR-010.
- The IEEE Learning Technology Standards Committee (LTSC) (http://ltsc.ieee.org/) is chartered by the IEEE Computer Society Standards Activity Board to develop accredited technical standards, recommended practices, and guides for learning technology. The Standard for Information Technology - Learning Technology - Competency Definitions (IEEE, 2003), (Mairtin, 2003) defines a universally acceptable Competency Definition model to allow the creation, exchange and reuse of Competency Definition in applications such as learning management systems, competency or skill gap analysis, learner and other competency profiles, and so on.
- The IMS Global Learning Consortium (http://www.imsproject.org/) develops and promotes the adoption of open technical specifications for interoperable learning technology. The scope for IMS specifications (IMS, 2003a, 2003b), broadly defined as "distributed learning", includes both online and off-line settings, taking place synchronously (real-time) or asynchronously. This means that the learning contexts benefiting from IMS specifications include Internet-specific environments (such as Web-based course management systems) as well as learning situations that involve off-line electronic resources (such as a learner accessing learning resources on a CD-ROM). The learn-

ers may be in a traditional educational environment (school or university), in a corporate or government training setting, or at home. IMS has undertaken a broad scope of work. They gather requirements through meetings, focus groups, and other sources around the globe to establish the critical aspects of interoperability in the learning markets. Based on these requirements, they develop draft specifications outlining the way software must be built in order to meet the requirements. In all cases, the specifications are being developed to support international needs

- The Advanced Distributed Learning (ADL) (http://www.adlnet.org/) initiative, sponsored by the Office of the Secretary of Defense (OSD), is a collaborative effort between government, industry and academia to establish a new distributed learning environment that permits the interoperability of learning tools and course content on a global scale. The following are several technologies the ADL initiative is currently pursuing:

  - Repository systems provide key infrastructure for the development, storage, management, discovery, and delivery of all types of electronic content.
  - Game-based learning is an e-learning approach that focuses on design and "fun".
  - Simulations are examples of real-life situations that provide the user with incident response decision-making opportunities.
  - Intelligent Tutoring Systems. "Intelligent" in the context of Intelligent Tutoring Sys tems (ITS) refers to the specific functionalities that are the goals of ITS development.
  - Performance Aiding (also called Performance Support) is one of the approaches being used to support the transformation. Improved human user-centered design of equipment, replacing the human role through automation as well as new technology for job performance are examples of the transformational tools under investigation to bridge gap between training, skills, and performance.
  - Sharable Content Object Reference Model (SCORM) was developed as a way to integrate and connect the work of these organizations in support of the DoD's Advanced Distributed Learning (ADL) initiative. The SCORM is a collection of specifications adapted from multiple sources to provide a comprehensive suite of e-learning capabilities that enable interoperability, accessibility and reusability of Web-based learning (Foix and Zavando, 2002).

- The Center for Educational Technology Interoperability Standards (CETIS) (http://www.cetis.ac.uk) represents UK higher and further education on international educational standards initiatives. These include the IMS Global Learning Consortium; CEN/ISSS, a European standardization; the IEEE, the international standards body now with a subcommittee for learning technology; the ISO, the International Standards Organization, now addressing learning technology standards.
- The ARIADNE Foundation (http://www.ariadne-eu.org/) was created to exploit and further develop the results of the ARIADNE and ARIADNE II European projects. These projects created tools and methodologies for producing, managing, and reusing computer-based pedagogical elements and telematics-supported training curricula. The project's concepts and tools were validated in various academic and corporate sites across Europe.
- Promoting Multimedia Access to Education and Training in European Society PROMETEUS (http://www.prometeus.org). The output from the Special Interest Groups within PROMETEUS could be in the form of guidelines, best practice handbooks, recommendations to standards bodies, or recommendations to national and international policy makers. The objectives of PROMETEUS are:

  1. To improve the effectiveness of the cooperation between education and training authorities and establishments, users of learning technologies, service and content providers, and producers within Europe
  2. To foster the development of common European and international standards for digital multimedia learning content and services

3. To give a global dimension to their cooperation, and to have open and effective dialogues on issues relating to learning technologies policy with policy makers in other regions of the world, while upholding Europe's cultural interests and specificities
4. To consider that the way to achieve these goals is by following certain common guidelines organizing their future cooperation
5. To consider that these guidelines should be based upon analysis of the needs expressed by users of the information and communication technologies in the education and training sectors.

In summary, the main and common goal for these standards is the reuse and interoperability of the educational contents between different systems and platforms.

There are many e-learning tools used in different context and platforms but, in general, Web-based training is the tread for the training process in many institutions.

Software tools used in Web-based learning are ranked by function:

1. **Authoring Tools:** Essentially, multimedia creation tools.
2. **Real-Time Virtual Classrooms:** A software product or suite that facilitates the synchronous, real-time delivery of content or interaction by the Web.
3. **Learning Management Systems (LMS):** Enterprise software used to manage learning activities through the ability to catalog, register, deliver, and track learners and learning. Within the learning management systems category, there are at least three subsets of tools:
   a. **Course Management Systems (CMS):** Software that manages media assets, documents, and Web pages for delivery and maintenance of traditional Web sites. It generally consists of functions including content manager, asynchronous collaboration tool, and learning record-keeper.
   b. **Enterprise Learning Management Systems (LMS):** Provides teams of developers with a platform for content organisation and delivery for a varied kind of content.
   c. **Learning Content Management Systems (LCMS):** A multi-user enterprise software that allows organisations to author, store, assemble, customise, and maintain learning content in the form of reusable learning objects.

Currently, there are many platforms and tools for e-learning (Table 1) (Foix and Zavando, 2002). In (EduTools, 2004) there is an interesting comparative study of e-learning tools from the functional point of view ranking by:

1. Communication tools: discussion forums, file exchange, online notes/journal, internal e-mail, real-time chat, video services, and whiteboard.
2. Productivity tools: bookmarks, calendar/progress review, orientation/help, searching within a course, work off-line/synchronize.
3. Student involvement tools: group work, self-assessment, student community building, student portfolios.
4. Administration tools: authentication, registration, course authorization, hosted services.
5. Course delivery tools: automated testing and scoring, course management, instructor helpdesk, student tracking.
6. Curriculum design: accessibility compliance, content sharing/reuse, course templates, curriculum management, customized look, instructional design, instructional standards compliance.

Some common features implemented for these tools can be summarized as:

- Security, authentication, firewall, and so on
- Client browser requirement
- Enrollment: online registration
- Searching
- Groupwork
- Tracking for student and courses
- Facilitating interface customization
- Asynchronous delivery and course management system
- Student information systems
- Enterprise resource planning systems: calendar and progress review

Table 1. E-learning tools

Supplier	Product
Blackboard (a.k.a. Courseinfo)	Blackboard http://products.blackboard.com
CBM Technologies	TEDS http://www.teds.com/
Université catholique de Louvain	Claroline : Open Source e-Learning http://www.claroline.net/
Docent	Docent Enterprise Learning Management Server, Training Partner http://www.docent.com/
Geometrix Systems	Training partners http://www.training-partners.com
IBM Mindspan Solutions/Lotus Software	LearningSpace www.learningspace.org/
IMC (information multimedia communication)	CLIX (Corporate Learning and Information eXchange)
Integrity eLearning	WBT Manager http://www.ielearning.com/wbt/index.cfm
IntraLearn Software	IntraLearn http://www.intralearn.com/
Knowledge Planet	KPLearning Management System http://www.knowledgeplanet.com/products/kp_learning.html
LearnFrame	Pinnacle Learning Management system http://www.learnframe.com/solutions/pinnacle/
Martin Dougiamas	Moodle- Modular Object-Oriented Dynamic Learning Environment http://moodle.org/
Pathlore Software	Pathlore Learning Management System http://www.pathlore.com/products_services/lms.html
Plateau Systems	Enterprise Learning Management System(ELMS) http://www.plateau.com/products/
Prometheus	Prometheus http://www.prometheus.com/product/
Saba Software	Saba Enterprise Learning Suite http://www.saba.com/english/products/
Technomedia	Sigal http://www.technomedia.ca/website/English/sigal_desc.htm
WBT Systems	TopClass http://www.wbtsystems.com/products
WebCT	WebCT Campus Edition http://www.webct.com/

- Discussion boards for student and community interest groups
- Structure and graphics design based on templates
- An online courses common catalog.
- Tools for abilities and competition evaluation.
- Learning evaluation systems: self-assessment, automated testing and scoring
- Materials or learning objects libraries
- Resources integration for the administration of knowledge
- Organizational information
- Individualized reports
- Spaces for knowledge and collaboration: discussion forums, file exchange, internal e-mail, chat, student community building, student portfolios, hosted services
- Content sharing and reuse
- Course templates

The e-learning process is not only for educational institutions; actually, more institutions use e-learning systems to train their employees

Non-educational institutions have specific needs and priorities in e-learning. The financial entities prioritize the speed in which courses can be designed for the delivery of new products and services. Courses for training in financial entities are characterized by:

- Speed in the generation: the delivery of a new product or service requires efficiency in the multimedia resources integration.
- Uniformity in design: the graphic designs are similar for each specific entity (color, logos, etc.).
- Specificity: each course is designed for a particular entity which does not want to exchange contents with other institutions.
- The contents belong to a specific domain and their structure is predetermined.

- The evaluation processes are simple.
- The information is restricted to the employees concerned.

With these characteristics, the use of complex e-learning tools is not a wise decision from the cost/benefit point of view.

This is the opposite of the standardization needs such as those outlined by the European Union regarding the e-learning program. This program and those of other of educational entities require a wide scope, variety of styles, designs, and mainly capacity to exchange contents (European Union, 2003).

THE E-FACTORY PROJECT

As explained before the e-factory project consists of two parts: the development of the factory tool, and a description of the results obtained using the factory tool.

Factory tool

Factory was proposed as the last of a chain of solutions the financial institution in question considered for the training process. So once e-learning was selected as the training option and the virtual campus (c@mpus) developed, the e-factory pilot project got started (1999). The first step was to develop the factory tool whose main features had to answer the following goals:

1. Factory had to be portable so that it could be easily installed in any personal computer in the financial entity. To achieve this, the use of JAVA code was decided on because the virtual Java machine can be executed in any personal computer.
2. Factory had to facilitate the development of the courses, minimizing time and cost of development. Therefore, factory was endowed with a set of modules which covers all the necessities of a course. The factory user can easily and quickly select not only the contents of the course but also its structure (for instance, the course must be structured in lessons, sections, and paragraphs), style (for instance, the background must be blue, all the course material must include the financial entity's logo and an exit button), and exercises. These imply a module able to generate structures, styles, and exercises as well as the correspondent modules to read structures, styles, exercises, and contents.
3. The factory had to generate courses completely ready to be published in the selected internet virtual campus. Therefore, factory generated HTML and XML courses.

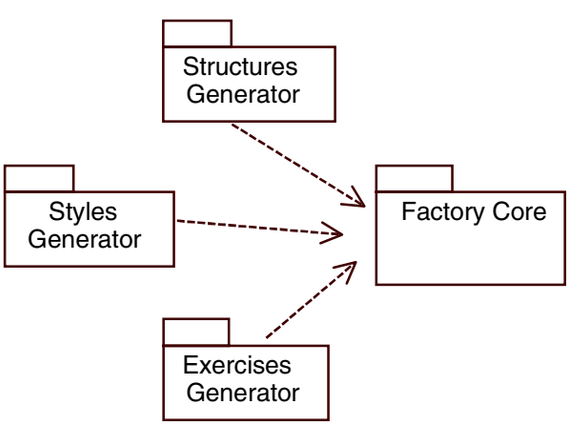
As a result of the above-mentioned goals, Figure 1 shows a main packages UML diagram to represent an overall view of the factory tool. The solution adopted for the factory tool allows:

- Easy inclusion of tracking sentences.
- Visualization of the courses using navigators that understand HTML and also include a module able to generate the same course in XML code, which can be interpreted by new generation navigators.
- Endowing semantic content to the courses using XML. The same course can be used by different students and, depending on their level, the contents shown will be different.
- Easy inclusion of new packages.

A brief description of each package follows:

· Structures generator package. This package must allow the development of different kinds of structures for the courses. Sometimes the same structure is used for courses of the same level;

Figure 1. Factory tool main pakages



with this package, the structure is generated once and reused on different courses.

- Styles generator package. This package must allow the development of different styles for the courses. Sometimes an organization has a corporate style that this must be used in all the courses during the same year. After that, they change some icon or image in the general style of the course. With the use of this package, one style is generated once and reused on all the courses. The style normally includes the general appearance of the course.
- Exercises generator package. This package generates exercises to be included on the courses. The kinds of exercises developed with this package are: drag and drop, join with an arrow, test, and simulation.
- Factory core package. This is the part of the system in charge of generating the web course gathering the contents and exercises, using a selected style and structure. This package has been broken down into in smaller ones (see Figure 2). These are explained below:

  - **Course tracking package.** This is the part of the system that includes code sentences in the courses generated to allow student tracking once the courses are allocated in a specific virtual campus.
  - **Structures reader package.** This is the part of the system that applies previously generated structures to a course to be developed.
  - **Styles reader package.** This is the part of the system that applies previously generated styles to a course to be developed.
  - **Exercises reader package.** This is the part of the system that includes previously generated exercises in the course under development.
  - **Contents reader package.** This is the part of the system that gathers contents for the course under development. These may be: text, multimedia elements, images and complex contents that are html code with JavaScript, etc.
  - **XML code generator package.** Once all the material is gathered, this part of the system generates the XML code corresponding to the course. The XML format was selected because it allows the development of different levels using the same course.
  - **HTML code generator package.** Since not all the client web browsers understand XML, we decided to endow this package with the functionality of translating the courses to HTML. In this case, the semantic potential of XML is lost.

Figure 2. Deep view of factory core package

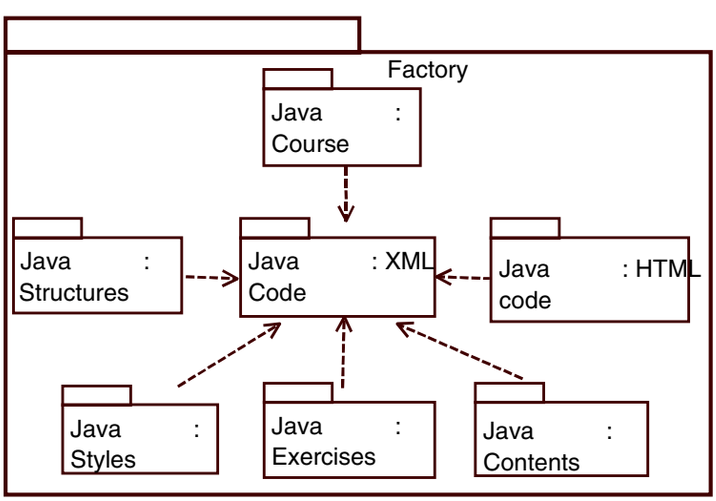


Figure 3. Factory general interface

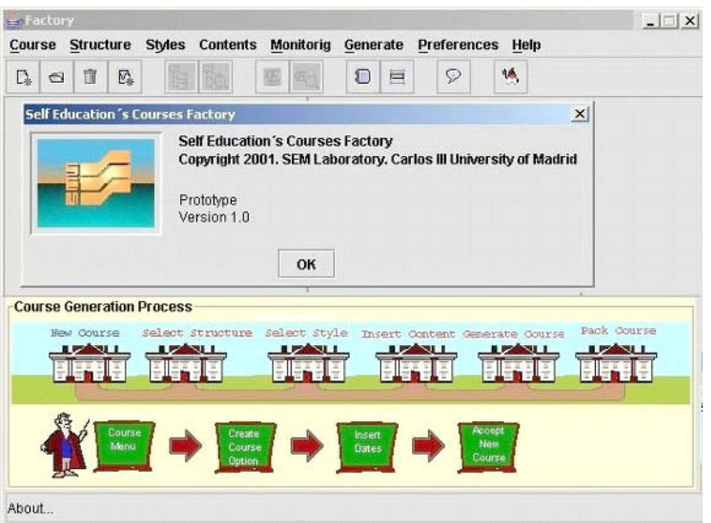
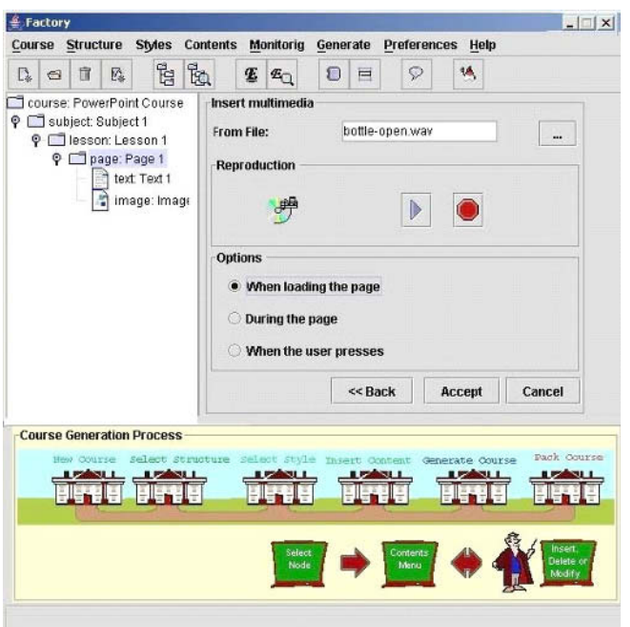


Figure 4. Factory form for multimedia resources integration



The Factory tool was developed with JAVA technology and used SQL server as database. The communication with the database was implemented using the bridge JDBC-ODBC. The course is generated in HTML or XML format. The main structure of the interface of the application can be seen in Figure 3.

The course structure allows the insertion of subjects and each subject can have one or more lessons. Each lesson can be structured in pages and each page can have information in the form of text, images, multimedia, or complex contents that represents complex html code and several kind of multimedia resources (see Figure 4).

Figure 5. Factory course generator

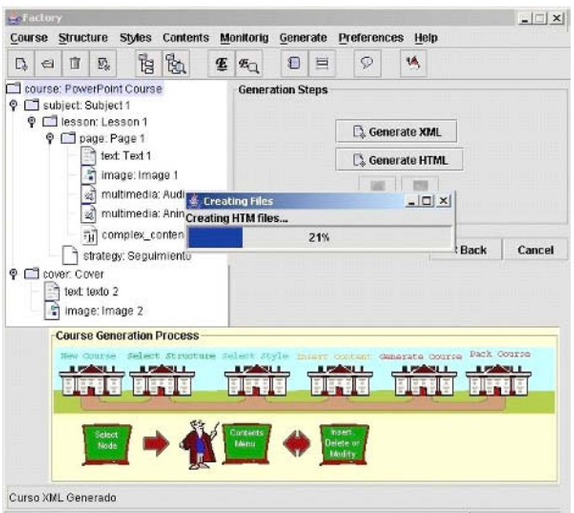


Table 2. Data comparison

	Without Factory			With Factory		
Course name	**Course Hours**	**Development Time (Hours)**	**Cost (euros)**	**Course Hours**	**Development Time (Hours)**	**Cost (euros)**
Leasing	15	560	14300	15	210	12000
House Credit	12	420	13300	15	210	12000
Credit Cards	15	560	15000	15	280	16000
Business on line	25	350	36100	20	210	12000
Oral Communication	9	280	12110	15	70	4000
Time Management	9	280	12110	15	70	4000
Meeting Management	8	280	12000	15	70	4000
Advanced Excel	15	350	12500	20	140	8000
EURO	10	280	12000	10	70	4000
Internet	10	280	12000	20	140	8000

Factory has a course generator process as a final process for the course publication. The product for this process is the HTML or XML pages (see Figure 5).

Factory allows a lot of interesting functionality that cannot be illustrated due to the extension of the paper.

The last part of the e-factory project was the use of the factory tool to compare cost and development time with those of the outsourced courses.

Table 2 shows the cost and time comparison between course development with and without the factory tool. As the courses developed without the factory tool were outsourced, we do not know exactly the kind of tool used in their development.

As can be seen above, if we take the leasing course as an example, it was developed in 560 hours when outsourced, as opposed to 210 hours using the factory tool. In relation to the cost, the leasing course cost 14,300 euros when outsourced while it costs 12,000 euros when it was developed using the factory tool, which included the cost to train the people in the use of the factory tool. So the benefits for the financial entity are significant and important.

CONCLUSION

The main e-learning problem in a financial institution arose when developing the course for a new financial product since, if three months were needed, the urgent training these courses demanded was lost. To solve this problem a factory tool allowed the development of new courses within a few weeks. This tool facilitated the rapid gathering and integration of contents in the courses.

The e-factory project has obtained excellent results in many aspects:

- The factory tool is able to develop courses in HTML and XML format of different levels of complexity. It can be easily enhanced, is portable because it is written in Java, and can be used by people with little computer science knowledge. It also covers the expectations of a research tool in the sense that the factory is developed using the latest tendencies in the area of new technologies and e-learning.
- The whole e-factory project has achieved all the proposed goals. From the data gathered it can be seen that the financial institution has notably reduced the cost in training its employees. It is also expected that the impact of new financial products on the market will produce an increment in the benefits of the financial institution.

FUTURE TRENDS

Although the factory was developed to solve the problems of training in financial institutions, in fact it would be a useful tool in any kind of institution with training needs. Usually, the contents of the courses are ready and, using factory, the teacher can easily prepare the lessons for the students in HTML or XML format without any knowledge of specific Web tools.

REFERENCES

Ahmad, H., Udin, Z.M., & Yusoff, R.Z. (2001). Integrated process design for e-learning: A case study. *Proceedings of the Sixth International Conference on Computer Supported Cooperative Work in Design (IEEE Cat. No.01EX472)*, (pp. 488-491).

AICC (1995). AICC Courseware Technology Subcommittee. *Guidelines for CBT Courseware Interchange,* October 31.

AICC (1997). AICC Courseware Technology Subcommittee. *Distance Learning Technology for Aviation Training,* June 24.

Comisión de las comunidades europas. (2004). Propuesta de decisión del parlamento europeo y del consejo por la que se establece un programa integrado de acción en el ámbito del aprendizaje permanente. 2004/153. Online *http://www.guiafc.com/documentos/2004-com-474.pdf*

Coppola, N.W. & Myre, R. (2002). Corporate software training: Is Web-based training as effective as instructor-led training? *IEEE Transactions on Professional Communication, 45*(3), 170-86.

EduTools (2004). Developed by WCET (the Western Cooperative for Educational Telecommunications) and from a grant from the William and Flora Hewlett Foundation. Online *http://www.edutools.info/course/compare/all.jsp*

E-learnframe (2004). Glossary of e-Learning Terms. Online *http://www.learnframe.com/*

Foix, C. & Zavando, S. (2002). Estándares e-learning: Estado del Arte Versión: 1.0, Centro de Tecnologias de Información, October 7.

Hodgkinson, M. & Holland, J. (2002). Collaborating on the development of technology enabled distance learning: A case study. *Innovations in Education and Training International, 39*(2), 89-94.

IEEE (2002). IEEE Computer Society Sponsored by the Learning Technology Standards Committee; *IEEE Std 1484.12.1™ 2002; IEEE Standard for Learning Object Metadata*, September 6.

IEEE (2003). IEEE Computer Society Sponsored by the Learning Technology Standards Committee; *IEEE Standard for Learning Technology Systems Architecture (LTSA), std1484.1TM,* Approved June 12.

IMS (2003a). IMS Abstract Framework: Applications, Services & Components v1.0. C. Smythe (Ed.), *IMS Global Learning Consortium, Inc.*, July.

IMS (2003b). IMS Abstract Framework: Glossary v1.0. C.Smythe, (Ed.), *IMS Global Learning Consortium, Inc.,* July.

LCT (2004). Language Coaching and Translations, Mag. Margit Waidmayr, e-Learning Glossary. Online *http://www.lct-waidmayr.at/e_glossary.htm*

Mairtin, S., O'Droma1, I.G., & McDonnell, F. (2003). Architectural and functional design and evaluation of e-learning VUIS based on the proposed IEEE LTSA reference model. *Internet and Higher Education, 6,* 263-276.

Presby, L. (2002). E-learning on the college campus: A help or hindrance to students learning objectives: A case study. *Information Management, 15*(3-4), 17, 19-21.

Seufert, S. (2002). *E-learning business models: Framework and best practice examples. Cases on worldwide e-commerce: Theory in action.* Hershey, PA: Idea Group Publishing.

Stark, C., Schmidt, K.J., Shafer, L., & Crawford, M. (2002). Creating e-learning programs: A comparison of two programs. *32nd ASEE/IEEE Frontiers in Education Conference, T4E-1,* November 6-9, Boston.

USD (2004). The University of South Dakota, Glossary of library and Internet terms. Online *http://www.usd.edu/library/instruction/glossary.shtml*

WBEC (2000). Report of the Web-based education commission to the president and the congress of the United States. *The power of the Internet for learning: Moving from promise to practice.* December.

KEY TERMS

ADL/SCORM ADLNet (Advanced Distributed Learning Network): An initiative sponsored by the US federal government to "accelerate large-scale development of dynamic and cost-effective learning software and to stimulate an efficient market for these products in order to meet the education and training needs of the military and the nation's workforce of the future." As part of this objective, ADL produce SCORM (Sharable Content Object Reference Model), a specification for reusable learning content. Outside the defence sector, SCORM is being adopted by a number of training and education vendors as a useful standard for learning content.By working with industry and academia, the Department of Defense (DoD) is promoting collaboration in the development and adoption of tools, specifications, guidelines, policies and prototypes that meet these functional requirements:

- Accessible from multiple remote locations through the use of meta-data and packaging standards
- Adaptable by tailoring instruction to the individual and organizational needs
- Affordable by increasing learning efficiency and productivity while reducing time and costs
- Durable across revisions of operating systems and software
- Interoperable across multiple tools and platforms
- Reusable through the design, management and distribution of tools and learning content across multiple applications

AICC (Aviation Industry CBT [Computer-Based Training] Committee): An international association that develops guidelines for the aviation industry in the development, delivery, and evaluation of CBT and related training technologies. The objectives of the AICC are to:

- Assist airplane operators in development of guidelines which promote the economic and effective implementation of computer-based training (CBT).
- Develop guidelines to enable interoperability
- Provide an open forum for the discussion of CBT (and other) training technologies

Authoring Tool: A software application or program used by trainers and instructional designers to create e-learning courseware. Types of authoring tools include instructionally focused authoring tools, web authoring and programming tools, template-focused authoring tools, knowledge capture systems, and text and file creation tools.

Case Study: A scenario used to illustrate the application of a learning concept. May be either factual or hypothetical.

Courseware: Any type of instructional or educational course delivered via a software program or over the Internet.

E-Learning Process: A sequence of steps or activities performed for learning purpose and using of technology to manage, design, deliver, select, transact, coach, support, and extend learning.

IEEE LTSC (Learning Technologies Standards Committee): Consists of working groups that develop technical standards in approximately 20 different areas of information technology for learning, education, and training. Their aim is to facilitate the development, use, maintenance, and interoperation of educational resources. LTSC has been chartered by the IEEE Computer Society Standards Activity Board. The IEEE is a leading authority in technical areas, including computer engineering.It is intended to satisfy the following objectives:

- Provide a standardized data model for reusable competency definition records that can be exchanged or reused in one or more compatible systems
- Reconcile various existing and emerging data models into a widely acceptable model
- Provide a standardized way to identify the type and precision of a competency definition
- Provide a unique identifier as the means to unambiguously reference are usable competency definition regardless of the setting in which this competency definition is stored, found, retrieved, or used
- Provide a standardized data model for additional information about a competency definition, such as a title, description, and source, compatible with other emerging learning asset metadata standards

Learning Management Systems (LMS): Enterprise software used to manage learning activities through the ability to catalog, register, deliver, and track learners and learning.

Training: A process that aims to improve knowledge, skills, attitudes, and/or behaviors in a person to accomplish a specific job task or goal. Training is often focused on business needs and driven by time-critical business skills and knowledge, and its goal is often to improve performance. See also Teaching and Learning.

Web Site: A set of files stored on the World Wide Web and viewed with a browser such as Internet Explorer or Netscape Navigator. A Web site may consist of one or more Web pages.

Supporting Online Communities with Technological Infrastructures

Laura Anna Ripamonti
Università degli Studi di Milano, Italy

INTRODUCTION

A lot of experiences with online communities (AOL, CompuServe, The WELL, Listserv and so forth) pre-date the Web, and some researchers have suggested that *"the origins of online communities were very close to the counter-cultural movements and alternative ways of life emerging in the aftermath of the 1960s"* (Castells, 2001, p. 53). The FreeNets movement, which emerged mainly in the United States (U.S.) and Canada in the second half of the 1980s, was basically aimed at providing citizens with free access to the Internet and providing content free from any form of control. In that framework, both Community and Civic Networks emerged, which are very nearly the same but for emphasis on the empowerment of the proximate community (Carroll & Rosson, 2003), on the "sense of community" and on the promotion of "citizens' participation in community affairs" (Schuler, 2001). FreeNets, Community and Civic Networks also shared features such as bottom-up development and, especially at their beginning, the use of Bulletin Board System (BBS) technologies (De Cindio & Ripamonti, 2004).

Due to the skyrocketing Web development, this type of community gradually evolved, giving birth to communities that cross the boundaries of organizations, countries, age groups and profit and non-profit organizations, and becoming *"a mainstream fixture for focused files, information and knowledge exchanges"* (Terra, 2003, p. 212). As a result, the term "online community" (in all its slightly different declinations: virtual community, Web community, network community, etc.) is currently used to define a wide range of social interactions taking place mainly on the Internet, generating a certain confusion since, as Preece (2000) points out, *"superficially, the term 'online community' isn't hard to understand, yet it is slippery to define"* (p. 9). The concept of online community seems to cover the whole rich panorama that has flourished, starting from the effective and intuitive characterization given by Rheingold (1994):

> *virtual communities are cultural aggregations that emerge when people bump into each other often enough in cyberspace. A virtual community is a group of people who may or may not meet one another face to face, and who exchange words and ideas through the mediation of computer bulletin boards and networks. (pp. 57-80)*

This early broad definition gave rise to debate (e.g., Jones, 1997, Wilbur, 1997, Levy, 1995), since it looks quite weak for distinguishing actual online communities from other kinds of Internet-based social aggregations. To further increase the confusion, too often—especially among computer scientists and scholars—the description of the software supporting online communities has been used as a "shortcut" for defining online communities, implicitly assuming that its appropriate use is the basis for building and maintaining an online community. Obviously, forums, mailing lists, chats, MUDs and so forth may support (physical) communities through the Internet, enriching their possibilities and extending their frontiers, since these communication technologies may favour the rising of (online) communities by scratch. However, setting up a discussion group, a mailing list or a MUD does not necessarily give birth to an online community, as experience has shown.

On the other hand, it has often been said that the technological part of the work related to community building is the "easy one," while careful and detailed planning of the system that regulates social interaction is absolutely fundamental in establishing successful and long-lived online communities.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Both positions have something to say, but we argue that a participatory approach, grounded in multidisciplinary studies and considering socio-organizational and technical aspects together, is indispensable for designing successful online communities (Ripamonti, 2003).

A GUIDING LIGHT IN A SEA OF "TECHNOLOGICAL FOG"

In spite of the different forms online communities may assume and the consequent different impacts they can have from a business point of view, they all share several common characteristics that can be used as a "guiding light" in designing appropriate technical infrastructure to support their development.

Among the major shared characteristics, for example, is that they all are socio-technical structures that place varying degrees of emphasis on belonging to communities of practice (CoP; see Wenger, McDermott & Snyder, 2002). Moreover, when they are built within one or more organizations, they also involve organizational aspects. They all imply knowledge sharing and collective thinking, a strong sense of belonging and mutual trust, a typical life cycle (Kim, 2000) and so forth. Besides these common factors, a number of minor characteristics can better describe any single community and explain the differences among them. For example, an online community may be large and quite loose, while another small and tightly knit, and so on.

The above considerations may sound quite "alarming" when analyzing the characteristics of the commercial technological platforms that support online communities since, at the moment, no technological solution seems to be a perfect fit for the needs of community. Both Preece (2000) and Wenger (2001) note that online communities are too often described in terms of "features," while – on the contrary – it is necessary first to define the critical success factors for community building and only then to find the "right" technological solution. Perhaps the most complete empirical map of these technologies has been traced by Wenger (2001): His analysis focuses on online communities of practice, but can be easily generalized—with some "tuning"—to online communities, since their definition can be seen as a generalisation of the CoPs' one, where several aspects have been "loosened," such as the strong stress on learning aspects. For this purpose, Wenger has produced an interesting model for classifying existing products that organizes community-oriented technological platforms on a map in relation to eight dimensions (*ongoing integration of work and knowledge, team work, social structure, discussion, fleeting interactions, instruction, knowledge exchange, documents handling*) significantly related to the main focus of the product, which can be combined in pairs representing the different aspects of the social life of knowledge. Applying his model, Wenger stresses that—at the moment—no ideal solution exists, ever for communities built within large business organizations.

DESIGN GUIDELINES FOR ONLINE COMMUNITY ENVIRONMENTS

Technology Does Matter

A general-purpose scheme to develop community-based technological environments, as already stated, should be based on the definition of a set of critical success factors affecting community development. To this extent, works from Wenger (2001), Wenger, McDermott and Snyder (2002), Kim (2000) and Terra (2003) can be of help; they list critical design aspects of three slightly different types of community (respectively, an online CoP, a general-purpose online community and a community for Knowledge Management), among which, as easily predictable, the existence of several common areas is immediately perceivable. These include: *attention to users' profiles, membership and roles; attention to the quality of the contents (use of netiquette, creation of value added); design for growth; presence and visibility (communications plans, promotion of events, etc.)*.

Nevertheless, Wenger is the only one to explore deeply how technological aspects affect critical success factors, with a special emphasis on the platforms aimed at supporting online activity (see Table 1). Thus, in designing a technological infrastructure, these factors should be constantly considered as a term of comparison for correctly translating needs into technological features.

Table 1. The 13 fundamental factors of successful CoPs that can be affected by technology (according to Wenger, 2001)

CRITICAL SUCCESS FACTOR TO BUILD COP/KM COMMUNITIES		
Factors	**Technological implications**	
1. Presence and visibility	• pointers • community directories • push distributions (newsletter, reminders …) • member directories	• who is doing what • presence awareness • instant messaging • virtual coffee
2. Rhythm	• community calendar • reminders • synchronization of calendars • synchronous events	• invitations • minutes of events quickly available • hot topics
3. Variety of interactions	• Asynchronous: • E-mail and discussion boards • document check-out/version control	• Synchronous: • lectures and large meetings • application sharing • Web tours
4. Efficiency of involvement	• integration with work systems • personalized knowledge/application portals • subscriptions	• tours of new activity • content filtering and ordering • archiving of interactions
5. Short-term value	• mechanisms for asking questions • FAQs • Databases of answers	• Intelligent access to experts • Help forums • Brainstorming facilities
6. Long-term value	• Repositories for artifacts • Taxonomies • Search mechanisms	• Discussing and updating a learning agenda • Spaces for practice-development projects
7. Connections to the world	• News • External events announcements • Directory of external experts	• Links to other sites • Library of references
8. Personal identity	• Profiles • Synchronizing profiles across communities with multiple view • Reputation and ranking	• Preferences • Personal history • Private places
9. Communal identity	• Having and furnishing a communal place • Public access to community's "source documents"	• News about the community and success stories • Distinctive look and feel • Community public presence
10. Belongings and relationships	• Personal profiles • Supporting private interactions and interpersonal relationships	• Conversations online to help the "shyest" • Chat moderators
11. Complex boundaries	• Differential access rights • Lurking facilities • Public areas and restricted ones	• Subspaces • Nested features
12. Evolution maturation and integration	• Low initial investment (money) to have a "tentative" commitment • Enough features for future development	• Flexibility in configuration • Ongoing reflection, assessment and redirection
13. Active community building	• Logs and statistics • Polling and voting • Assessment tools and surveys	• Administrative help and reminders • Switches and policy enforcements algorithms • Health indicators

A PROACTIVE PARTICIPATORY METHODOLOGY IS ESSENTIAL

Building a Web environment based on communities implies both adopting a multidisciplinary approach and considering critical success factors for community development. In other words, once the critical success factor set is defined, a methodological framework is needed to provide sufficiently powerful tools to investigate how they should be deployed in the specific case and to understand the technological, social and organizational implications of such a multifaceted problem. Hence, it sounds quite difficult to adopt a totally theoretical approach, partly because no one can judge better than the community members themselves how "good" whichever solution a well-meaning developer may invent. This means it is necessary to couple practical action with theory and to foster a strong, direct interaction with the specific subjects that will participate in the community.

A good way to meet these challenges is through the adoption of the Action Research (AR) methodology, which offers *"an approach to research that aims both at taking action and creating knowledge or theory about that action"* (Coughlan & Coghlan, 2002, p.220). As Westbrook (1995) points out, AR could overcome some deficiencies in traditional research methods, since it also has broad relevance for practitioners and applies to unstructured or integrative issues.

Nevertheless, AR alone is unable to guarantee the appropriate strong emphasis on the participatory aspect of community environment design, which is necessary to guarantee that critical success factors are correctly translated into technical functionalities. To reach this goal we suggest coupling AR with Community-Centered Development (CCD) (Preece, 2000). By and large, CCD is a participatory process that involves members of a specific community, and whose goal is to develop a social and technical infrastructure capable of sustaining an online community throughout its whole life cycle.

The CCD approach is about *online community building* (starting from the design phase and covering the whole community life cycle): It assumes that both technological and sociability issues need to be carefully planned to develop a successful online community, and it guides developers through several fundamental steps. Its characteristics are such that one can quite easily couple it with AR, obtaining a resulting "methodological" sum that offers a more suitable research framework than either of the two could if used "stand alone."

The CCD borrows ideas from multiple multidisciplinary sources, but most of them derive from theories developed in the computer science field. Fundamentally, CCD derives principles from *User-centred design* (e.g., Norman, 1986) and *Interaction design* (e.g., Preece, Rogers, Sharp, Benyon, Holland & Carey, 1994; Preece, Rogers & Sharp, 2001; Kreitzberg, 1998; Shneiderman, 1998), which focus on users rather than simply on technologies; *Contextual inquiry*, which underlines the relevance of understanding the user context (e.g.,

Figure 1. AR cycle (according to Coughlan & Coghlan, 2002)

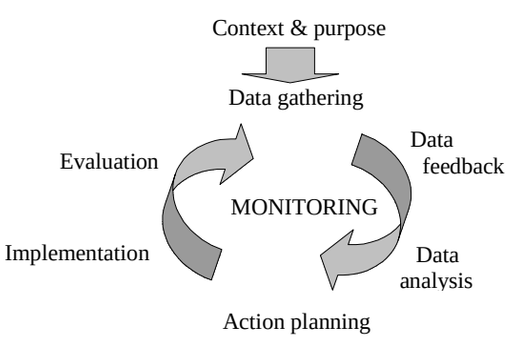


Figure 2. Usability and sociability issues in CCD (Preece, 2000)

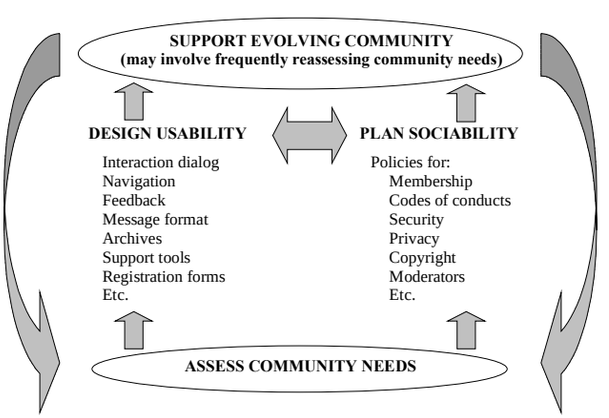


Figure 3. CCD steps (according to Preece, 2000)

Community needs assessment and user task analysis
Selecting technology and planning sociability
Designing, implementing and testing prototypes
Refining and testing sociability and usability
Welcoming and nurturing the community
EVALUATION

Beyer & Holtzblatt, 1998); and *Participatory design*, which emphasizes the necessity of a strong user and community participation in the design process (e.g., Mumford, 1983;Muller, 1992; Schuler & Namioka, 1993). All these methods stress that it should be users' needs that to some extent shape the technological infrastructure, and not the other way around, as happens quite often, unfortunately. Actually, as Preece (2000) points out:

The relationship between the design of any artifact, the way people use it, and how it both affects and is affected by social norms is complex. Software is no exception. And when the software is intended to support social interaction, understanding this relationship becomes very important. ... Developing communities involves technical considerations. Unlike most other software, for successful online communities, sociability is also essential. (pp. 204-205)

Hence, CCD must focus on community needs before any decision about technology and social planning has been made; and therefore, the issues of *usability* (appropriateness of the software design for the community members' tasks and needs) and of *sociability* (appropriateness of the social policies and plans for the social interaction) are both addressed by the CCD.

FUTURE TRENDS

Wenger and Snyder (2000a) and Terra (2003) describe different ways in which online communities can create value. These include: helping drive strategies, starting new lines of business, solving problems quickly, transferring best practices, developing professional skills, helping companies recruit and retain talent, winning new businesses more quickly, better serving existing clients, developing stronger relationships with clients, facilitating the integration of acquired companies or during post-merger phases, reducing cross-functional and cross-location cultural barriers, improving organizations' social capital, reducing costs and playing a significant role in merger and acquisition operations. As a result, despite the fact that few organizations (of any size) have so far paid close attention to online communities, according to Gartner Group (2001), by the end of 2004, more than 50% of Fortune 500 enterprises will formally support at least online CoPs. The growing interest of the business world for online communities, interwoven with rapidly evolving technologies, will presumably lead to the rapid development of new, more effective, integrated tools for supporting online relationships, leading to convergence among the different technologies and devices (e.g., wireless, Web, handheld PC, virtual reality, etc.) used to connect to shared virtual spaces. For

this reason, increased attention should be paid to design; otherwise, users are unlikely to populate technologically perfect worlds.

CONCLUSION

In the last decade, the Web has had a skyrocketing development. As a consequence, from the early 1980s, social-focused, online communities have budded a phenomenon—with business implications as well—that grows rapidly, flourishing into the wide variety of online "communities" we now come across every day (MUDs, MOOs, discussion groups, newsgroups, community networks, etc.). Nonetheless, they all share a well-defined set of basic characteristics that can be analyzed to determine several factors critical for successful community building. To this end, technological, social and organizational aspects have to be jointly considered (with varying emphasis according to the specific context) to develop any online community intended to live and prosper. In particular, the "technological part" is not at all the simplest issue in community building, as has been said at times, since technical choices may have heavy drawbacks for the creation and development of relationships among community members. Hence, it is important to verify that the specific set of critical success factors has been correctly translated into technological features. A good checklist that couples critical factors with their technological counterpart was drawn up by Wenger (2001). Furthermore, to provide a proper handling of such a multifaceted problem as understanding the implications of the development process of an online community, a strong and appropriate methodological approach has to be adopted. A good methodological framework can be achieved by coupling AR with CCD.

REFERENCES

Beniger, J. (1987). Personalization of mass media and the growth of pseudo-community. *Communication Research, 14*(3), 352-371.

Beyer, H., & Holtzblatt, K. (1998). *Contextual design: Defining customer-centered systems*. San Francisco: Morgan Kaufmann.

Carroll, J.M., & Rosson, M.B. (2003). *A trajectory for community networks*. In H. Sawhney (Ed.), *Special issue ICTs and community networking. The Information Society International Journal, 19*(5), 395-406.

Castells, M. (2001). *The Internet galaxy: Reflections on the Internet, business and society*. Oxford: Oxford University Press.

Coughlan, P., & Coghlan, D. (2002) Action research for operations management. *International Journal of Operation & Production Management Emerald, 22*(2), 220-240.

De Cindio, F., & Ripamonti, L.A. (2004). Nature and roles for community networks in the information society. To appear in P. Day (Ed.), *Special Issue Community Informatics, AI & Society Journal.*

Gartner Group. (2001). *Communities: Broad-reaching business value*. GartnerGroup Publication N. COM -13-9032, July 3.

Jones, S. (1997). The Internet and its social landscape. In S.G. Jones (Ed.), *Virtual culture: Identity and communication in cybersociety* (pp. 7-35). Thousand Oaks, CA: Sage.

Kim, A.J. (2000). *Community building on the Web*. Berkeley, CA: Peachpit Press.

Kreitzberg, C. (1998). *The LUCID Design Framework (Logical User-Centered Interaction Design).* Princeton: Cognetics Corp.

Levy, P. (1995). *Quest-ce que le virtuel*. Paris: La Dècouverte.

Muller, M.J. (1992). Retrospective on a year of participatory design using the PICTIVE technique. *Proceedings of CHI '92 Human Factors in Computing Systems*, May 3-7, Monterey, California.

Mumford, E. (1983). *Designing participatively*. Manchester: Manchester Business School.

Norman, D.A. (1986). Cognitive engineering. In D. Norman & S. Draper (Eds.), *User-centered systems design*. Hillsdale: Lawrence Erlbaum.

Preece, J. (2000). *Online communities: Designing usability, supporting sociability*. Chichester, UK: John Wiley & Sons.

Preece, J., Rogers, Y., & Sharp, H. (2001). *Interaction design: Beyond human-computer interaction*. New York: John Wiley & Sons.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., & Carey T. (1994), *Human-computer interaction*. Wokingham: Addison-Wesley.

Rheingold, H. (1994). A slice of life in my virtual community. In L.H. Harasim (Ed.), *Global networks: Computers and international communication*. Cambridge, MA: MIT Press.

Ripamonti, L.A. (2003). *Online communities for knowledge sharing in SMEs*. PhD Thesis, D.I.Co. University of Milan, Italy.

Schuler, D. (2001). Cultivating society's civic intelligence: patterns for a new 'world brain.' *Journal of Information, Communication and Society, 4*(2).

Schuler, D., & Namioka, A. (Eds.). (1993). *Participatory design: principles and practices*. Hillsdale: Erlbaum.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (3rd ed.). Reading, MA: Addison-Wesley.

Terra, J.C. (2003). Twelve lessons to develop and sustain online knowledge communities. *Proceedings of IADIS International Conference e-Society 2003,* June 3-6, Lisbon, Portugal

Wenger, E. (2001). *Supporting communities of practice: a survey of community-oriented technologies*. Draft version 1.3. Retrieved March 2001 from *www.ewenger.com/ewbooks.html*

Wenger, E.C., & Snyder, W.M. (2000a). Communities of practice: The organizational frontier. *Harvard Business Review,* Jan/Feb, 139-145.

Wenger, E.C., & Snyder, W.M. (2000b). Learning in communities. *LINE Zine*. Retrieved from *www.linezine.com*

Wenger, E., McDermott, R., & Snyder, W.M. (2002). *Cultivating communities of practice: A guide to managing knowledge*. MA: Harvard Business School Press.

Westbrook, R. (1995). Action research, a new paradigm for research in production and operations management. *International Journal of Operations & Production Management, 15*(12), 6-20.

Wilbur, S. (1997). An archaeology of cyberspace. Virtuality, community, identity. In D. Porter (Ed.), *Internet culture*. New York: Routledge.

KEY TERMS

Action Research: An approach to research that aims both at taking action and creating knowledge of theory about that action (Coughlan & Coghlan, 2002).

Community-Centered Development (CCD): A participatory process that involves members of a specific community, whose goal is to develop a social and technical infrastructure able to sustain an online community throughout its whole life cycle (e.g., Preece, 2000).

Community Network/Civic Network (CN): The FreeNets movement, which emerged mainly the United States and Canada in the second half of the 1980s, was basically aimed at providing citizens with free access to the Internet (in a moment in which access was expensive) and granting content completely free from any attempt of control. In the framework of the FreeNets movement emerged both Community and Civic Networks, which are very much the same but for emphasizing the empowerment of the proximate community (see Carroll & Rosson, 2003) and of the "sense of community," thanks to the possibilities offered by the ICT, the role they can play in promoting "citizens participation in community affaires" (Schuler, 2001) and in reducing distances between citizens and (local) government bodies. FreeNets and Community and Civic Networks also shared features such as bottom-up development and, at their beginning, the use of the BBS technology.

Community of Practice (CoP): A group of people sharing expertise and passion about a topic and whom interact on an ongoing basis to further their learning in this domain (e.g., Wenger & Snyder, LINE zine 2000; Wenger, McDermott & Snyder, 2002).

Online Community: People who interact through an ICT-based communication environment, recognize a minimum common goal that holds them together, share one or more domains of knowledge and shared practices, and define implicit or explicit policies for regulating their interactions.

Online Community of Practice: A community of practice interacting mainly (or only) online (see Ripamonti, 2003).

Participatory Design: A methodology that emphasizes the necessity of a strong user participation in the design of a technological infrastructure (e.g., Mumford, 1983; Muller, 1992; Schuler & Namioka, 1993).

Pseudocommunity: Those ephemeral experiences defined as "communities" during the Internet economy boom period and that disappeared immediately after, since they didn't achieve the network of interactions and relationships that grants the sense of belonging in "true" communities (e.g., Beninger, 1987; Jones, 1997).

Teletranslation

Minako O'Hagan
Dublin City University, Ireland

INTRODUCTION: A BRIEF OVERVIEW

The translation industry developed in response to the need to assimilate or disseminate information across different languages where the text in one natural language had to be converted into another. Strictly speaking, "translation" deals with written text, whereas "interpreting" handles spoken dialogues, although the latter is often subsumed in the former in common terminology. Translation services are called upon when the sender and the receiver of the message do not share a common language. Although translation is not a new area of professional practice, the industry as a whole has been transformed due to the recent advancement of information and communication technology (ICT). While computer applications ranging from text processing tools to Machine Translation (MT) have affected the production of translation, telecommunications technology has changed the operational dynamics of translation services. The development of electronic networks and computer applications to translation gave rise to the concept of teletranslation (O'Hagan, 1996). Furthermore, ICT is creating new types of content which becomes subject to translation, driving new types of language support. Teletranslation can be characterized by increased reliance on ICT in terms of transmission, storage, and processing of translation in contrast to conventional translation which dealt with off-line text distributed in print. Today modern translation operators are using the Web as a customer-interface and also as a platform to interconnect a team of translators and tools in different locations, dealing with the new types of text embedded in various electronic forms.

Prior to the 1980s, physical transportation systems underpinned the translation business by providing the link among translator, customer, and translation agency that acted as an intermediary. As such translation services were constrained by physical distance, thus operating primarily as a regional business. However, the arrival of fax machines allowed them to become less location-bound. The nature of translation work primarily being text-based and asynchronous (i.e., translation does not usually have to be done instantaneously) suited telework mode in which translators receive and return text at a distance. Fax machines facilitated telework by allowing freelance translators to work on text remotely without incurring additional delivery delays. During the 1980s and into the early 1990s, text transmission via fax was gradually replaced by use of modems with text transmitted directly from computer to computer. This provided the advantage of text arriving in electronic form with flexibility for further processing, as compared with a faxed hard copy. Into the 1990s, the power of computer networks began to see translation businesses operate internationally, linking translators and customers worldwide.

In the mid-1990s, the Internet began to permeate the physical national borders representing different languages and cultures and make information available online in multitudes of languages, in turn driving the need for translation. For example, a user stumbling across a foreign language Web site would seek indicative translation on the spot without having to leave the computer terminal, preferably at little or no cost since the value of the information is uncertain. Tapping into such needs, online MT became commonly available to translate Web sites or search engine results on the fly and provide the user with the gist of the content in a requested language. MT found a niche market which was not suited to human translation (HT) in view of cost, speed, and logistics. In this way, the Internet boosted the demand for MT applications (Tanaka, 1999). In the meantime, businesses that started to leverage the Internet to reach customers on a global basis realized the need for their Web sites to be available in different languages. This led to a new type of language service called Web localization which became the fastest growing area within the translation sector in the late 1990s (Lockwood, 1999). In this way, ICT has significantly altered the landscape of translation which was originally developed on the basis of translating non-electronic text in non-real-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

time for off-line consumption in the least computer-aided environment. Teletranslation can be defined as new modes of translation dealing primarily with digital content in computer-aided translation (CAT) environments, operating over global communications networks.

BACKGROUND: ASSOCIATED DEVELOPMENTS OF LOCALIZATION

ICT has had the effect of generating new types of translation work. The prime examples are computer software products whose needs for language support led to the development of the new sector called localization, which has developed since the early 1980s (Esselink, 2000). The localization industry came into existence to meet the demands of the opening international market for computer products; software and hardware needing to be adapted to the requirements of local customers. In addition to print-based translation of manuals and packaging, software localization integrates translation into the software engineering process and thus departs from the original nature of translation work. During the 1990s localization evolved continuously in response to new demands driven by globalization of a diverse range of products and services. Localization Industry Standards Association (LISA) lists fourteen sectors as localization service users, including medical/pharmaceutical and telecommunications in addition to software (Fry, 2003). Although localization incorporates translation, the former developed largely independently of the existing translation model with its own workflow and processes. This reflects the different nature of the task involved in localization as a whole; it has often been linked more closely to software engineering than to translation (O'Hagan, 2004). However, more recently, the two have begun to come together, reflecting the new areas of translation demand arising from digital media, in turn requiring localization beyond conventional translation. Today, language support is increasingly placed in the context of so called GILT (Globalization, Internationalization, Localization and Translation)—a term recently used by the localization industry. Localization, once considered as an esoteric sector within translation, is now rapidly being mainstreamed, subsuming translation. Teletranslation is therefore significantly influenced by the localization model which developed on the direct link between language and ICT.

The traditional concern of translation scholars, the translation process performed by human translators unaided by ICT, does not adequately account for the transformation of the industry outlined above (Sager, 1993). The dynamic nature of today's translation market and the role of translation in the context of GILT are calling for a framework in which the impact of ICT can be considered.

DISCUSSION: TELETRANSLATION AND TRANSLATION-MEDIATED COMMUNICATION

Teletranslation attempts to capture the significance of the driving force of ICT behind the transformation of conventional translation. As an analytical framework for taking account of the changes driven by ICT, an approach based on translation-mediated communication (TMC) was introduced (O'Hagan & Ashworth, 2002). While any communication mediated by translation can be regarded as TMC, the term is used specifically in analogy with computer-mediated communication (CMC). TMC is therefore concerned more with communication taking place in a CMC mode than with conventional modes un-mediated by computer. This approach incorporates the examination of translation as communication. Various translation scholars (Bell, 1991; Nida & Taber, 1969) have applied the communication model to translation, not always without criticism (Robinson, 2003), because of the model's primary focus on the mechanistic assessment in accuracy of transmission. In TMC, the correct transmission of the signal does not always mean the message intended by the sender was adequately conveyed through translation; it is concerned with the role of technology in terms of its impact on the sender, the receiver and the message as well as the translator in a qualitative, rather than a quantitative manner.

In the context of teletranslation, the communication model provides a useful basis whereby TMC can be examined as a process in which the sender of the message in the source language communicates with its receiver via translation which converts the message into the target language. Furthermore, the factors relating to ICT such as how the message is transmit-

ted, stored and processed can be incorporated. In this manner, the proposed analytical framework takes into account such aspects as the change in the nature of the message and the relationship among the sender, the receiver and the translator as well as their respective roles in the translation production process, all of which are affected by ICT applications. O'Hagan (2004) demonstrates how the framework based on TMC can be applied to analyzing the process involved in localization of Web sites. Recently emerged practices such as localization of various digital contents cannot be adequately explained solely in terms of equivalence between the source and the target texts as localization often involves extensive adaptation, depending on the requirement of the target audience. Furthermore, it also involves non-textual elements such as images, icons, the page layout, colours, and so on which encroach upon the field of intercultural communication.

The following sections elaborate on two of the significant changes affecting TMC, which emerged from localization practice.

Internationalization

Today's product globalization often involves the pre-localization process called internationalization. This process facilitates an efficient localization process by incorporating into the original products factors affecting the ensuing localization, including translation. This necessitates the sender of the message considering the requirements of the receiver at the beginning of the product development cycle in order to minimize likely problems during the subsequent localization. Here, the sender designs the source content with localizability (and translatability) in mind. In this way, the relationship between the sender and the receiver becomes much closer than with conventional translation where the message is typically created by the sender without particular consideration for the receiver. The process involves technical adjustments such as non-ASCII (American Standard Code for Information Interchange) character set support for Asian languages, changes in date/time conventions, making allowance for text expansion where the target text equivalent is longer in length than the source text, and so on. The suitability of certain non-textual components such as images and icons also needs to be considered in relation to the target audience requirements. Internationalization facilitates all the possible subsequent changes without having to re-engineer the entire product when it is localized for a specific market.

One of the hallmarks of conventional translation has been the way in which it was treated as an afterthought in relation to the creation of the original content to be translated. In contrast, a significant aspect of internationalization which is now beginning to be widely applied to product globalization is the idea of involving translators in the knowledge creation chain. This relates to foundational multilinguality discussed in Joscelyne (2000) as practiced by certain international organizations whereby the source text is authored by a multilingual team often writing in their non-native languages. This prompts the in-house translation section to become closely involved in filtering the source content by having regard to translation production. As a result, translators actively contribute to the organization's knowledge products instead of conducting translation as an isolated downstream activity. This is the essence of internationalization in which the consideration of translatability influences the creation of the source content. While internationalization is a relatively well explored topic from software engineering perspectives (Kano, 1995), developing an international design for software products is complex and the practical difficulties in absorbing all problems up-front are also well articulated (O'Sullivan, Denner, & Galligan, 2004). For example, similar observations will apply to emerging language support practices for video games and multimedia content on DVDs (O'Hagan, 2004; O'Hagan & Mangiron, 2004). Further research is needed to establish a holistic internationalization model that is applicable to a wider range of content and includes insights from linguistic and cultural perspectives.

Collaborative Translation on Global Networks

Another significant development is collaborative translation undertaken by a virtual team supported by electronic networks. Conventionally, a translation job would be handled by a single translator from beginning to end. Time pressure, increased job complexity and ensuing requirements for different types of skills for completing projects often demand team-based work. This is common particularly for large-scale localization projects but also increas-

ingly adopted for other types of translation work. To maximize the advantage of collaborative translation, certain CAT tools also need to be sharable on the network. For example, translation memory (TM) which stores previous translations is now frequently used in network mode whereby allowing translators working on the same project to leverage the previous translations or terminology database created by other translators. This contributes to standardization of terminology and phraseology among different translators working on different parts of the same project. Production efficiency and product quality for teletranslation heavily depend on how effectively the human talents and computer systems are deployed in a collaborative environment. In this way, future developments of workflow systems and CAT tools to facilitate collaborative work significantly affects teletranslation.

FUTURE PROSPECTS

The main rationale behind teletranslation was to accommodate the new dimensions of translation practice being rapidly developed in response to new demands coming from, and capabilities afforded by, ICT. Teletranslation continues to evolve as new technology produces new content which requires language support. Many new translation demands today are coming from digital media including video games and audiovisual content on DVDs. What distinguishes these from traditional translation text is that the former require translation of multimedia components which in turn need to be embedded within the digital environment of the original content. This calls for localization which integrates the linguistic and engineering processes. Conventional translation typically dealt with static written text in the main, but teletranslation needs to respond to dynamic content which may be in hypertext form and incorporate multimedia involving speech, text, and images. The development of new digital content is also requiring the integration of currently separate branches of translation such as screen translation. For example, creating different language versions of video games involves processes similar to software localization and screen translation in the form of subtitling and dubbing (Mangiron, 2004). Screen translation which involves language transfer in audiovisual content in turn is beginning to be influenced by some of the approaches adopted in localization workflow, for example, in using templates for the production of multilingual subtitles for DVDs (Georgakopoulou, 2004).

In order to explore further into future possibilities, a series of teletranslation experiments were conducted with a range of Internet-based chat environments (O'Hagan, 2000; O'Hagan, 2002; O'Hagan & Ashworth, 2002). The platforms chosen included chat rooms with avatars (computer graphic representations of participants), allowing a varying degree of control on nonverbal cues. On the basis of these experiments, a number of hypotheses were formulated regarding the future implications for teletranslation. The first was that the conventional division between translation and interpretation would become less distinct. For example, a CMC mode such as text chat involves the real-time exchange of written text which takes place in interactive mode. The language support for this mode needs to incorporate synchronous production of translation whereby the text is translated in real-time. Translation by humans in this environment thus entails a new skill combining interpreting, dealing with synchronous communication, with translating, which involves written text.

Another hypothesis made was that nonverbal communication would become a significant component subject to language support. In conventional interpretation practice, nonverbal cues are something implicitly taken into account. For example, interpreters do not directly reproduce nonverbal traits displayed by the speaker, such as facial expressions or mannerisms although these may be implicitly processed by the interpreter, influencing the interpreter's rendition of the speech. Desktop virtual reality (VR) environments, often prevalent in games, are beginning to provide an interactive communication space including nonverbal communication cues. The degree of sophistication of avatars to indicate certain facial expressions and body movements varies. These environments point to the potential in future to allow more finely tuned nonverbal cues according to each participant's cultural background. For the language support providers, this will enable language assistance to extend beyond the verbalized message to include a wider communication context which involves nonverbal cues. This in turn will require a transformation of some of the

basic assumptions of traditional translation and interpretation whereby subjecting nonverbal cues to explicit, rather than implicit processing.

These hypotheses were further reinforced through thought experiments using HyperReality as an example of an advanced VR environment. HyperReality represents one of a number of similar areas of research being undertaken in laboratories all over the world (Tiffin & Terashima, 2002); the technology in its final form seeks to merge VR into Real Reality (Ibid). As such, it provided an ideal context in which the future potential of teletranslation could be explored. Interactions in HyperReality take place in a specially designated space called coaction fields, which in turn provide all-digital highly customizable environments (Ibid). Such advanced digital communication environments suggest a possibility to create a virtual multilingual communication space in which a person could communicate in spoken or written form, synchronously or asynchronously, in a specified language while using a full range of nonverbal communication cues that can also be localized for the benefit of the communicating parties. Teletranslation in such an environment could be designed to provide transparent communication assistance, drawing on a global network of HT and MT resources to facilitate both verbal and nonverbal communication.

One future scenario created is a Virtual Polyglot Space (VPS) (O'Hagan, 2000; O'Hagan & Ashworth, 2002), which allows anybody who enters the space to communicate multilingually through ubiquitous language support. Such services are envisaged to be provided by a team of translators and MT connected via global networks. One of the major features of VPS is its ability to customize contextual nonverbal cues in an appropriate form in relation to the cultural context suitable for the communicating party. For example, for a Japanese participant, all the other participants and elements in the given virtual reality environment can be designed to display nonverbal cues according to Japanese conventions, whereas an Irish user in the same VPS will experience the conventions to be based on Irish customs. In designing such environments, internationalization and collaborative translation will come into play.

These remain as speculative possibilities for now. However, the direction of ICT and its impact on teletranslation suggests that communications environments in future could be furnished with such ubiquitous multimodal language support.

CONCLUSION

The concept of teletranslation was originally proposed to accommodate emerging language support practices which did not fit the conventional translation model. Today there are many translation operators offering their services using the term teletranslation, evidence of the need to differentiate their services from those of conventional translation. The drive behind the need for a new conceptualization of translation comes from the advancement of ICT. Teletranslation will continue to evolve and its future will be driven by the direction of ICT.

REFERENCES

Bell, R.T. (1991). *Translation and translating: Theory and practice*. London: Longman.

Esselink, B. (2000). *A practical guide to localization*. Amsterdam; Philadelphia: John Benjamins Publishing.

Fry, D. (2003) *The localization primer* (2nd ed.), revised by Arle Lommel. Féchy: Localization Industry Standards Association (LISA).

Georgakopoulou, Y. (2004). DVD subtitling: A search for the Holy Grail? A paper given at the *International Conference on Audiovisual Translation*. In So Many Words:University of London, February 6-7, 2004.

Joscelyne, A. (2000). The role of translation in an international organization, in R.C. Sprung (Ed.), *Translating into success*, (pp. 81-95). Amsterdam; Philadelphia: John Benjamins Publishing.

Kano, N. (1995). *Developing international software for Windows 95 and Windows NT*. Washington: Microsoft.

Lockwood, R. (1999). You snooze, you lose. *Language International, 11*(4), 12-14.

Mangiron, C. (2004). Localizing final fantasy: Bringing fantasy to reality. *LISA Newsletter*, XIII, 1.3.

Nida, E.A. & Taber, C. (1969). *The theory and practice of translation*. Leiden: E.J. Brill.

O'Hagan, M. (1996). *The coming industry of teletranslation*. Clevedon: Multilingual Matters.

O'Hagan, M. (2000). *Hypertransterpretation in HyperReality*. Doctoral Thesis. Wellington, Victoria University of Wellington.

O'Hagan, M. (2002). HyperTranslation. In Tiffin & Terashima (Eds.), *HyperReality: Paradigm for the third millennium*. London: Routledge.

O'Hagan, M. (2004). Conceptualizing the future of translation with localization. *The International Journal of Localization*, *1*(1), 15-22.

O'Hagan, M. & Ashworth, D. (2002). *Translation-mediated communication in a digital world*. Clevedon: Multilingual Matters.

O'Hagan, M. & Mangiron, C. (2004). Games localization: When "Arigato" gets lost in translation. *Proceedings of New Zealand Game Developers Conference*, Dunedin, New Zealand, (pp. 57-62).

O'Sullivan, P., Denner, G.., & Galligan, N. (2004). Global software development: A non-English perspective. *The International Journal of Localization*, *1*(1), 29-39.

Robinson, D. (2003). *Performative linguistics: Speaking and translating as doing things with words.* London, Routledge.

Sager, J. (1993). *Language engineering and translation: Consequences of automation*. Amsterdam; Philadelphia: John Benjamins Publishing.

Tanaka, H. (1999). What should we do next for MT system development? *Proceedings of Machine Translation Summit VII 99*, Tokyo, Asia-Pacific Association for Machine Translation (AAMT), (pp. 3-8).

Tiffin, J. & Terashima, N. (Eds.). (2002). *HyperReality: Paradigm for the Third Millennium*. London: Routledge.

KEY TERMS

Chat: An interactive CMC model between two or more people who can enter text by typing on the keyboard, and have the entered text appear in real-time on the monitors of all participants. This can be done via speech. Internet Relay Chat (IRC) is an early example.

Computer-Mediated Communication (CMC): CMC was made widely popular by the Internet, which allows people to communicate in a variety of modes such as e-mail or chat. CMC in turn is affecting translation practice as more and more people communicate in a CMC mode across languages and require language support.

Globalization: The process of planning and implementing products and services so that they can be adapted to different local languages and cultures.

Internationalization: The process to facilitate localization in the context of globalization.

Interpretation: The process involved in converting the source speech in one natural language into the target speech in another language.

Localization: The process of adapting a product or service to a particular language, culture, and desired local "look and feel".

Machine Translation (MT): A computer program which translates text from one natural language into another.

Telework: Use of computers and telecommunications to enable people to work remotely away from the office. The substitution of telecommunications for transportation.

Translation: The process involved in converting the source text written in one natural language into the target text in another language.

Translation-Mediated Communication (TMC): Communication mediated by translation in CMC rather than in conventional communication modes.

Virtual Reality (VR): A technology which provides an interactive interface between human and computer that involves using multiple senses, typically sound, vision, and touch in the computer generated environment.

Virtual Team: A group of people working on the same project from different locations linked by computer networks.

Telework Information Security

Loreen Marie Butcher-Powell
Bloomsburg University of Pennsylvania, USA

INTRODUCTION

The sophistication of technology available to businesses as well as to homes has increased dramatically in the last 10 years. The speed of information exchange and the ease of use of computer software have become a major influence on the decision of businesses to allow unconventional working environments. As a result, telework has become an increasingly preferred option to working in the office (Manochehri & Pinkerton, 2002). In the early 1970s, Jack Nilles coined the word *telework*. Telework refers to an approved working arrangement whereby an employee—a teleworker—officially performs his or her assigned job tasks in a specified work area of his or her home on a regular basis (United States Department of Defense, 2002). According to the Communications Security Establishment's Telework Pilot Program (2002), telework has become a very important alternative work pattern, which allows employees to better manage their home life and work life in a complex society. Telework offers many advantages, including the following:

1. Substantial savings in physical facility-related costs, including rent, storage, and electricity;
2. Expanding labor pools without geographic restrictions (Hirsch, 2002; Mehlman, 2002; Motskula, 2001).

After the September 11, 2001, terrorist attacks on the United States of America, many corporations turned toward telework (Niles, 2001; United States Department of Defense, 2002). However, with the increased benefits afforded by teleworking, there are increased security risks, including viruses and data tampering (Atwood, 2004; Hirsch, 2002; Motskula, 2001; Quirk, 2002; Rubens, 2004).

BACKGROUND

In order to understand the importance of securing the infrastructure for telework, the scope of the term *security* from a telework infrastructure perspective must be defined. The term *security* leads one to investigate the survivability of a network or related assets to an attack (Allen, 2001). This is the key to the integrity of the data resident on the network system and alludes to the flexibility of network assets to cope with internal and external intrusions and corruption (Allen, 2001). Further, the survivability of a network and its associated economics need to be assured, regardless of the transmission or storage media (Landwehr, Bull, McDermott, & Choi, 1994). However, the economics affiliated with security issues are not limited to the system level, but rather it can extend within the confines of the network infrastructure (Allen, 2001). According to Dhillion and Backhouse (2000), information systems security in a telework environment must address both the data and the changing organizational context in which data are interpreted and used.

Existing research has shown that in order to prevent sensitive data from being disclosed, modified, or made unavailable in transit between two endpoints, the communications link must be protected (Davis, 2001; Hercovitz, 1999). Threats that utilize or take advantage of the communications link can be wire tapping, replay attacks, man-in-the-middle attacks, war dialers, denial of service attacks, and buffer overflow attacks. Possible safeguards to help mitigate these threats include firewalls, authentication and access control measures, and virtual private networks (VPNs). However, VPNs have remained the most popular way to secure a telework infrastructure (Davis, 2001; DeSanctis et al., 1996).

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

VPNs use familiar wide area networking (WAN) technology and protocols. Generally, a client or workstation using WAN technologies sends a stream of encrypted point-to-point protocol (PPP) packets to a remote server or router. This same process occurs with VPNs, except instead of going across a dedicated line, the packets go across a tunnel over a shared network such as the Internet. VPNs allow teleworkers to gain remote access to a specific corporate network via the Internet by tunneling into the corporate intranet (Brown, 1999).

VPNs are a combination of tunneling and encryption algorithms that carry traffic over the Internet, a managed Internet Protocol (IP) network, or a service provider's backbone network (Stallings & Van Slyke, 1998). It provides encrypted tunnels through the Internet that permit off-sites to communicate securely. Tunnels provide a secure path for network applications. For example, if a teleworker wants to connect into the corporate network to access a company's intranet, this individual can dial into or connect to his or her local Internet service provider (ISP) and connect as though he or she were onsite. The teleworker can then initiate a tunnel request to the destination security server on the corporate network. The security server authenticates the user and creates the other end of the tunnel. Next, the teleworker sends data through the tunnel. Data are encrypted by the VPN software before being sent over the ISP or Internet connection. The destination security server receives the encrypted data and decrypts the packets. The security server forwards the decrypted data packets onto the corporate network. This same encryption process applies if any information is sent from the corporate network to the teleworker (Davis, 2001).

Traditionally, VPNs utilize both tunneling methods and encryption algorithms to carry traffic over the Internet. For example, network traffic reaches the VPN backbone using any combination of access technologies, including T-1, Frame Relay, Integrated Services Digital Network (ISDN), Asynchronous Transfer Mode (ATM), or simple dial-up access (Davis, 2001).

The most commonly accepted method for creating VPN tunnels is called Layer 2 Tunneling (L2T). L2T is created by encapsulating a network protocol such as IPX, NetBEUI, and AppleTalk inside the Point-to-Point Protocol (PPP), and then encapsulating the entire package inside a tunneling protocol. Traditionally, L2T VPN packets are sent toward the remote network to reach a tunnel-initiating device. The tunnel initiator will then communicate with a VPN terminator, or a tunnel switch, to agree on an encryption scheme. The tunnel initiator then encrypts the packet for security before transmitting to the terminator. The terminator then can decrypt the packet and deliver it to the appropriate destination on the network (Brown, 1999).

VPNs have provided a solid foundation for corporations to secure a telework infrastructure (DeSanctis et al., 1996). However, this is changing. The number of security breaches increases as the number of teleworkers increases (ITAC, 2004).

AN INCREASE IN SECURITY BREACHES OCCURRING

The number of teleworkers in the U.S. is increasing at unprecedented rates. According to the International Telework Association Council (ITAC) (2004), the American Interactive Consumer Survey revealed that 16.8 million Americans teleworked in 2001, and 23.5 million Americans teleworked in 2003. Cahners In-Stat, a leading provider of actionable research, has predicted that by the year 2005, more than 60% of the workforce will be considered teleworkers (DecisionOne, 2002).

As the number of teleworkers increased, the number of security breaches has also increased. A 2000 CSI/FBI Computer Crime and Security Survey revealed this marked increase in computer security incidents increased as a direct result of increase usage of the Internet. The survey showed the Internet as a frequent point of attack (POA) (Federal Bureau of Investigation/Computer Security Institute, 2001). Additional research also states that telework increases the risk of potential security threats, including viruses and data tampering occurring within any corporate network (Hirsch, 2002; Powell, 2002).

Companies are beginning to be concerned about the information security problems from teleworking. An American Telephone and Telegraph Company (AT&T) survey and white paper, in cooperation with the economist, stated that 49% of the 237 telework corporations surveyed have security concerns regarding the electronic communication of telework. The

paper further stated the need for additional research in securing the corporate infrastructures (AT&T, 2003).

Traditional information security methodological approaches focusing on the formal automated parts of the corporate infrastructure via corporate checklists, evaluations, and risks analysis have been employed by companies to revolve some security issues (Dhillion & Backhouse, 2000). However, many security breaches still remained. Hirsch (2002), Mehlman (2002), Motskula (2001), and Powell (2002) stated that the severity of security threats in a telework infrastructure is often related to the computer literacy of the teleworker accessing the network rather than the actual corporate network. For example, a white paper provided by Teleworker.org (2003) suggested that teleworkers using a high-speed Internet connection such as a cable modem or digital subscriber line (DSL) create an even higher security risk for telework infrastructures. Because high-speed Internet connections are always connected to the network, this increases the chance of the teleworker's computer being discovered by hackers running automated port scans.

Furthermore, an outside intruder may find it simpler to attack a less fortified teleworker's laptop that is logged on to the corporation network than to directly attack the corporate network itself (Hirsch, 2002). Not all teleworkers' laptops are protected at all times by the corporate network. Consequently, if a teleworker's laptop is not connected to the corporate network, it could be used for Web surfing, new software programs that are not related to work, old software reconfiguration, opening of e-mail attachments, and downloading of Internet files. Therefore, the laptop is not compliant with the corporate standards, thereby decreasing the effectiveness of any security software and increasing the risk of virus infection (Carlson, 2000).

Understanding the security risks of teleworkers is important in order to secure its telework infrastructure. The demand to secure telework access to corporate resources generates a need for the redesign and development of security models creating a new way to secure telework infrastructures (Atwood, 2004). The security models developed must support increasingly complex telework infrastructures and personnel. Hirsch (2002), Mehlman (2002), Motskula (2001), and Powell (2002) suggest that additional research regarding teleworker vulnerabilities is needed to help secure future telework infrastructures. They also assert that this type of research is critical, because without it, future telework infrastructures may result in corporate failure.

FUTURE TRENDS/RESEARCH

When considering telework infrastructure security assessment methodologies, companies are faced not only with the issue of how to adapt the traditional model to a telework environment, but also how to adapt technology to teleworkers. To address this challenge, research is being undertaken to secure end-users/teleworkers via modifying existing information security evaluation methodologies. For example, an existing and respected security framework developed by the Networked Systems Survivability (NSS) Program at the Software Engineering Institute (SEI) at Carnegie Mellon University, Pennsylvania, is the Operationally Critical Threat, Asset, and Vulnerability Evaluation ^{SM} (OCTAVE^{SM}) model. The OCTAVE ^{SM} model is a repeatable methodological approach for identifying and managing information security risks of actual threats, including disclosure of a critical asset, modification of a critical asset, loss or destruction of a critical asset, or interruption of access to a critical asset via an organizational self-assessment (Alberts, Behrens, Pethia, & Wilson, 1999).

The OCTAVE^{SM} model was designed "to be easily modified to meet the needs of many organizations" (Alberts & Dorofee, 2003, p.241), including telework corporations. This model currently is being studied, investigated, modified, and validated via a Delphi approach at the Bloomsburg University of Pennsylvania. A Delphi process was utilized to develop and validate a specific set of criteria necessary for the successful inclusion of end-user-/telework-based activities. The Delphi process modified OCTAVE ^{SM} model through 67 identified computer competencies and was administered to 437 technology coordinators, senior managers, and teleworkers from a large corporation located on the East coast of the U.S. After modifying the OCTAVE^{SM} model to include the identified set criteria for teleworkers, the OCTAVE^{SM} process began.

While this research is still in progress, preliminary results published at the time of this article (August 2004) indicate that the most important competency

that end users/teleworkers should develop is the ability to correctly utilize virus scanners and backup procedures on their computers (M=4.55, SD=0.87) and correctly use the firewalls' further protection (M=4.54, SD=0.94). In addition, the use of spyware and the understanding of operating system controls (M=4.15, SD=1.01) were considered important skills for end users/teleworkers to possess.

It is expected that further outcomes will result in a better-proposed corporate security plan that includes proper education and training for teleworkers on a yearly basis. However, it is important to understand that not all end users/teleworkers will yield the same results. Therefore, further research is still being conducted to develop a proposed security risk evaluation model that can be utilized by all similar-sized telework corporations.

CONCLUSION

Many corporate infrastructures feature high volumes of sensitive and confidential data relevant to internal and external transactions. Research has shown that telework connections put data at risk due to the potential for intrusion or eavesdropping resulting from the end user/teleworker's computer (Dhillion & Backhouse, 2000; Myersson, 2002). Every telework computer is susceptible to a variety of computer threats. Threats and risk assessments can be used to list the potential threats for a particular computer. Remote computers require protection; otherwise, sensitive data existing on the remote computer could be disclosed, modified, or made unavailable. Various threats to the remote computer itself include viruses, data modification, trojan horses, trap doors, sabotage, human error, and scavenging. To proactively develop a security evaluation plan that assesses the end user/teleworker's system and safeguards telework infrastructures, further research needs to be conducted on modifying existing security evaluation models to include the end user/teleworker.

ACKNOWLEDGEMENTS

Special permission to use the OCTAVE[SM] Method ©2004 by Carnegie Mellon University, Pennsylvania, in Loreen Butcher-Powell study was granted by the Software Engineering Institute.

REFERENCES

Alberts, C., Behrens, S.G., Pethia, R.D., & Wilson, W.R. (1999). Operationally critical threat, asset, and vulnerability evaluation (OCTAVE SM) framework, version 1.0. *(CMU/SEI-00-TR-017)*. Retrieved June 1, 2004, from *www.sei.cmu.edu/publications/documents/99.reports/99tr017/99tr017abstract.html*

Alberts, C., & Dorofee, A. (2003). *Managing information security risks: The OCTAVE[SM] aApproach*. Upper Saddle River, NJ: Addison-Wesley.

American Telephone and Telegraph Company (AT&T). (2003*). Remote working in the net-centric organization.* Retrieved June 27, 2004, from *http://www.business.att.com/content/whitepaper/remote_working_net-centric_org.pdf*

Atwood, S. (2004). Data protection: Protecting your remote office data using replication technologies. *Disaster Recovery Journal, 17*(2), 20.

Brown, S. (1999). *Implement virtual private networks.* New York: McGraw-Hill.

Carlson, P.A. (2000). Information technology and the emergence of a worker-centered organization. *ACM Journal of Computer Documentation*, *24*(4), 204-212.

Communications Security Establishment. (2002). *Government of Canada's, telework project (ITSPSR-14).* Retrieved October 1, 2003, from *http://www.csecst.gc.ca/en/documents/knowledge_centre/government_publications/itspsr/Telework Project_e.pdf*

Davis, C. (2001). *IPSEC: Securing VPNs*. New York: McGraw-Hill.

DecisionOne. (2002). *Creating an effective telework plan: What works, what doesn't, and why*. Retrieved June 1, 2004, from *http://www.decisionone.com/d1m/news/white_papers/white_paper_04.shtml*

DeSanctis, G., Jackson, B., Poole, M., & Dickson, G. (1996). Infrastructure for telework: Electronic communication at Texaco. In M. Igbaria (Ed.), *Pro-*

ceedings of the ACM SIGCPR/SIGBOT (pp. 94-102). New York: Association for Computing Machinery.

Dhillon, G. (2001). Violations and safeguards by trusted personnel and understanding related Information security concerns. *Computers & Security, 20*(2), 165-172.

Dhillon, G., & Backhouse, J. (2000, July). Information system security management in the new millennium. *Communications of ACM, 43*(7) 125-128.

Federal Bureau of Investigation/Computer Security Institute. (2001). *2000 CSI/FBI computer crime and security survey.* Retrieved July 13, 2004, from *http://www.gocsi.com*

Herscovitz, E. (1999). Secure virtual private networks: The future of data communications. *International Journal of Network Management, 12*(1), 213-220.

Hirsch, J. (2002). Telecommuting: Security policies and procedures for the "work-from-home" workforce. Retrieved June 1, 2004, from *http://www.teleworker.org/articles/telework_security.html*

International Telework Association & Council (ITAC). (2004). Retrieved March 18, 2005, from *http://www.telecommute.org/resources/abouttelework.htm*

Mehlman, B.P. (2002). Telework and the future of American competitiveness. Retrieved February 1, 2004, from *http://www.technology.gov/Speeches/BPM_020923_Telework.htm*

Motskula, P. (2001). *Securing teleworking as an ISP service*. Retrieved June 2, 2004, from *http://www.telework2001.fi/Motskula.rtf*

Myersson, J. (2002). Identifying enterprise network vulnerabilities. *International Journal of Networking Management, 12*(1), 135-144.

Niles, J. (2001). Telework and terrorism. Retrieved June 2, 2004, from *http://www.jala.com/world_trade_center.php*

Powell, L. (2002). Telework security: What users don't understand. *Proceedings from the Telebalt Conference 2002*, Vilnius, Lithuania.

Quirk, K.P. (2002). Telework in the information age. Retrieved June 2, 2004, from *http://www.accts.com/telework.htm*

Rubens, P. (2004). *What you need to nell neleworkers*. Retrieved June 1, 2004, from *http://networking.earthweb.com/netsecur/article.php/3306781*

Stallings, W., & Van Slyke, R. (1998). *Business data communications*. Upper Saddle River, NJ: Prentice-Hall Inc.

Teleworker.org. (2004). Retrieved June 1, 2004, from *http://www.teleworker.org/articles/telecommuting_security.html*

United States Department of Defense (n.d.). *Telework policy*. Retrieved June 1, 2004, from *http://www.telework.gov/policies/dodpolicy.asp#definitions*

KEY TERMS

ATM: Another form of packet switching in which fixed-size cells of 53 octets are used (Stallings & Van Slyke, 1998).

Authentication: The process of determining whether someone or something is who or what they declare to be. In private or public computer networks, authentication is commonly done through the use of logon passwords or digital certificates.

Automated Port Scan: An intruder sending a request to a host name or a range of IP addresses followed by a port number to see if any services, including file transfer protocol (FTP), telnet and hypertext transfer protocol (HTTP), are listening on that port. Automated port scans typically are carried out by hackers trying to gain large amounts of information about a particular network so that an attack can be planned.

Data Tampering: The threats of data being altered in unauthorized ways, either accidentally or intentionally.

Encryption: A combination of key length, key exchange mechanism, rate of key exchange, and key generation. Popular encryption algorithms are Kerberos, Data Encryption Standard (DES), and

Rivest, Shamir, and Adelman (RSA) Data Security (Power, 2000).

Firewall: A system or collection of systems that enforces an access control policy among networks. Generally, a firewall sits between the internal network and the outside network to block unauthorized traffic. For instance, when a user sends a message, the message goes through the firewall before going to the Internet. The corporate firewall typically sits between its internal network and the Internet communications environment. The corporate firewall serves as a gateway, blocking an unauthorized remote user from accessing the corporation's network, if the remote user does not have a valid IP address (Herscovitz, 1999). Because the firewall functions on the network and transport layers of the Open Systems Interconnection (OSI) Reference Model, remote users are able to authenticate to the firewall, receive an IP address that is valid for the local subnet, and then authenticate to the corporate server, as if they were physically connected to the in-house network.

Frame Relay: A form of packet switching based on the use of variable-length link layer frames.

ISDN: A telecommunication service that uses digital transmission and switching technology to support voice and digital communications.

Private Network: A specific organization's network (Dhillon & Moores, 2001).

Public Network: The Internet (Dhillon, 2001).

Remote User: A teleworker or individual who uses computer technology to work at home, on the road, or from other locations outside of the company's location (Nilles, 2001).

Risks Analysis: Techniques for providing a means of forecasting critical events. Risk analysis enables the identification of necessary controls that must be incorporated in a secure information system.

Scavenging: Attacking the physical access to a computer.

Security Evaluation: A systematic examination of an organization's technology base, including information systems, practices and procedures, administrative and internal controls, and physical layouts. Security evaluations identify security deficiencies such as outdated virus definitions and unauthorized access to information.

T-1 Access: Provides a data transmission rate of 1.544 megabits per second (Mbps).

Telework: An approved working arrangement between the employee and a company, whereby an employee officially performs his or her assigned job tasks in a specified work area of his or her home on a regular basis (United States Department of Defense, 2002).

Teleworker: An employee who officially performs his or her assigned job tasks in a specified work area of his or her home on a regular basis (Nilles, 2001).

Trap Doors: Pieces of code inserted into a program. Trap doors generally are used for the purpose of debugging or bypassing standard access control mechanisms.

Trojan Horse: A program that appears to perform only advertised tasks, but which is in reality performing further, often malicious, undocumented tasks in the background by the unauthorized user.

Tunneling: The encapsulating of protocol information for the transmission of data via the Internet to a private network.

Virtual Private Networks (VPNs): A private network that provides access to various locations via a public network carrier.

Viruses: Malicious programs that usually are transmitted by means of various types of files, including executable files. Viruses can shut down a PC and an entire network, delete files, and change files.

Wide Area Networks (WANs): The oldest type of network that extends to large geographical areas such as a state, nation, or the world.

Text-to-Speech Synthesis

Mahbubur R. Syed
Minnesota State University, USA

Shuvro Chakrobartty
Minnesota State University, USA

Robert J. Bignall
Monash University, Australia

INTRODUCTION TO SPEECH SYNTHESIS

Speech synthesis is the process of producing natural-sounding, highly intelligible synthetic speech simulated by a machine in such a way that it sounds as if it was produced by a human vocal system. A text-to-speech (TTS) synthesis system is a computer-based system where the input is text and the output is a simulated vocalization of that text. Before the 1970s, most speech synthesis was achieved with hardware, but this was costly and it proved impossible to properly simulate natural speech production. Since the 1970s, the use of computers has made the practical application of speech synthesis more feasible.

In principle, a TTS system is a two-step process (Figure 1) in which text is converted to its equivalent digital audio. A text and linguistic analysis module processes the input text to generate its phonetic equivalent and performs linguistic analysis to determine the prosodic characteristics of the text. A waveform generator then produces the synthesized speech (Carvalho, Trancoso, & Oliveira, 1998).

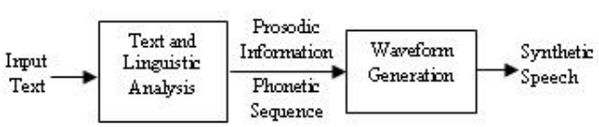
In the following sections, speech synthesis and its applications are described. Its historical development is outlined and the key challenges encountered by developers are summarized. A brief description is provided of some TTS synthesizers available at present.

APPLICATION OF SPEECH SYNTHESIS

Text-to-speech synthesis has potential applications in any domain in which speech is necessary or enhances communication with users. It has applications in education, telecommunications, consumer products, and a range of other areas. Imagine a TTS engine built into presentation software that not only shows a series of slides but also reads them to the audience. An important use of TTS synthesis is for the disabled, for example, to read an e-book or e-mail to a visually impaired person, or to read text typed by a deaf or vocally handicapped person. TTS synthesis becomes significant and useful if the audio information that needs to be communicated is too extensive to be stored, too expensive to prerecord, or if its recording is impossible because the application does not know ahead of time what output might be needed. For example, in a telecommunication application, a back-end TTS system may be able to read credit card information in real time over the phone to a customer. In such situations, TTS synthesis could offer a significant advantage over prerecorded audio sound if the information to be communicated is searched from a large database and cannot be anticipated in advance, or if the prerecording of audio information is not feasible.

Benefits of text-to-speech include reading dynamic text, conserving storage space, providing audible feedback, notifying a user of an event, proofreading documents, and so on. The usefulness of TTS synthesis in different application domains has at-

Figure 1. A two-step representation of a text-to-speech system



Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

tracted researchers to take up the challenge of developing natural sounding and intelligible TTS systems in many languages around the world (Carvalho et al., 1998; Möbius, 1999; Mukherjee, Rajput, Subramaniam, & Verma, 2000; Syed, Chakrabartty, & Bignall, 2004).

A BRIEF HISTORY OF SPEECH SYNTHESIS

Speech synthesis has gone through a long development process from the mechanical to the electrical to the electronic era. According to Schafer and Markel, one of the first electrical synthesizers was the Voder (Voice Operation Demonstrator), which attempted to produce connected speech and followed the principles of source-tract separation. The electrical networks in the device could be selected by finger-actuated keys, whose resonances were similar to those of individual spoken sounds. This speaking machine was demonstrated by trained operators at the World Fairs of 1939 (New York) and 1940 (San Francisco). To produce speech, the operators could play the device as if it was an organ or a piano, but it required that they undergo a year or so of training (Schafer & Markel, 1979).

"Before the 1980s, speech synthesis research was limited to large laboratories that could afford to invest the necessary time and money for hardware. In the mid-1980s, more laboratories and universities started to join in as the cost of the hardware fell. By the late '80s, purely software-based synthesizers that not only produced reasonable quality speech but could do so in near real time became feasible" (Black & Lenzo, 2003). Up to the present time, several software companies and research groups have developed a variety of mono- to multilingual TTS systems on a range of platforms. Some are described in a later section. A historical archive of audio clips produced by different TTS systems can be found at http://www.cs.indiana.edu/rhythmsp/ASA/Contents.html.

METHODS OF SPEECH SYNTHESIS

- **Formant Synthesis:** "Formant synthesis is a source filter method of speech production in which the vocal tract filter is constructed from a number of resonances similar to the formants of natural speech" (Donovan, 1996).
Formant synthesis uses resonators and filters to emulate the human vocal system. Formant synthesis provides an infinite number of sounds, enabling maximum flexibility in voice customization and it requires three formants to produce intelligible speech and up to five formants to produce high-quality speech. Each formant is usually modeled with a two-pole resonator that enables both the pole-pair formant frequency and its bandwidth to be specified (Donovan, 1996). Rule-based formant synthesis uses a set of rules to determine the parameters necessary to synthesize a desired utterance using a formant synthesizer (Allen, Hunnicutt, Klatt, Armstrong, & Pisoni, 1987). Typically, a fundamental voicing frequency (F0), three other formant frequencies (F1, F2, F3) and three formant amplitudes (A1, A2, A3) are used in this synthesis (Lemmetty, 1999).
- **Concatenative Synthesis:** Concatenative synthesis is done by connecting prerecorded natural utterances to produce intelligible and natural-sounding, arbitrary synthetic speech. The natural speech segments are selected and stored in an acoustic inventory. According to Lemmety, the selection of the optimal speech-unit length is one of the most important decisions, requiring a trade-off between longer and shorter units. High naturalness, fewer concatenation points, and good control of coarticulation are achieved with longer units, but the number of required units and the amount of memory needed are increased. Less memory is needed with shorter units, but sample collecting and labeling procedures become more difficult and complex (Lemmetty, 1999). The prerecorded elements of natural utterances cannot be whole words because of their large number, different forms, and constant new additions to the language. "In 1958, Wang and Peterson proposed the 'diphone' (the acoustic chunk from the middle of one phoneme to the middle of the next phoneme) since coarticulatory influences tend to be minimal at the acoustic center of a phoneme. With this more satisfactory unit, they estimated that as many as 8,000 diphones may

be necessary for a TTS system. However, current systems are able to function with an inventory of about 1,000 diphones" (http://www.mindspring.com/~ssshp/ssshp_cd/dk_758.htm). In TTS systems, the units used are usually phonemes, diphones, triphones, words, syllables, and demisyllables.

CHALLENGES IN SPEECH SYNTHESIS

According to Lemmetty, the generation of accurate prosody and pronunciation, and the correct handling of numerals, abbreviations, and acronyms are among several problems that arise in text preprocessing. "The expression of appropriate emotions and pronunciation of proper and foreign names is difficult and often very ambiguous. Speech synthesis with female and child voices has been found to be more difficult. A female voice typically has a pitch almost twice as high as a male voice, and with children, the pitch may be as much as 3 times as high. A higher fundamental frequency makes it more difficult to estimate formant frequency locations" (Lemmetty, 1999). The conversion of input text into a linguistic representation, usually called text-to-phonetic or grapheme-to-phoneme conversion, is complex and highly language dependent and conversion from text to speech can be divided into three main phases and is described below along with related challenges (Dutoit, 1997).

- **Text Preprocessing:** Text preprocessing includes various language-dependent problems that make it more complex. Several examples, as below, illustrated by Möbius and Lemmetty give a clear notion of the complexity of text processing. The spoken version of digits or numerals is context dependent. For example, in English, *1920* needs to be spoken as "nineteen-twenty" if a year is meant, or as "one thousand nine hundred and twenty" when it is a quantity. The spoken version of *1/4* may be expanded as "a quarter" if it is a fraction, or as "January the fourth" if a date is meant. Roman numerals have similar contextual problems. *Chapter III* should be expanded as "chapter 3," and *Henry III* as "Henry the Third." The text *I* may be either a pronoun or a number. Roman numerals may be also confused with some common abbreviations, such as *MC*. When expanding ordinals, the first three, namely *1st* as "first," *2nd* as "second," and *3rd* as "third" are exceptions. Numbers may also require some special form of expression. For example, *22* may be read as "double two" (for a telephone number), and *1-0* as "one love" (for scores in certain sports). There are contextual problems with expanding abbreviations; for example, *kg* may mean either "kilogram" or "kilograms." *Dr.* may be "doctor" or "drive," while *ft.* may be "fort," "foot," or "feet." Special characters and symbols such as $, %, &, /, -, and + may cause problems. In some situations, the word order must be changed. For example the text *$71.50* must be expanded as "seventy-one dollars and fifty cents," and *$100 million* as "one hundred million dollars," not "one hundred dollars million." The text *4-9* may be expanded as "four minus nine" or "four to nine." Some characters in Web sites or e-mail messages must be expanded using special rules. For example, the character @ is usually pronounced as "at," and e-mail messages may contain some strings, such as header information, that should be omitted (Black & Lenzo, 2003; Carvalho et al., 1998; Lemmetty, 1999; Möbius, 1999).
- **Pronunciation:** Another major task in the conversion process is to find the correct pronunciation for different contexts in the text. In some cases, the pronunciation does not correspond to its written spelling, as in *son*, *put*, *trough*, *though*, *thought*, and so on. Some words, called homographs, are spelled the same way but differ in meaning and usually in pronunciation; for example, the text *lives* is pronounced differently in the sentences "A cat has nine lives" and "The cat lives." Some words, such as *lead*, not only have different pronunciations when used as a verb or a noun, but may differ between two noun senses ("He lead a march" vs. "He followed a lead" vs. "He melted some lead"). With these kinds of words, some semantic information is necessary to achieve the correct pronunciation (Möbius, 1999; Möbius, Sproat, van Santen, & Olive, 1997).

- **Prosody:** Researchers agree that finding the correct intonation, stress, and duration (known as the prosodic or suprasegmental features) of speech to be generated from written text is perhaps the most challenging problem of all. The prosody of continuous speech depends on the meaning of the sentence, the speaker's characteristics, intentions, and emotions, the audience being addressed, and so forth. Information about such features and their dynamic changes during speech are not available in written text. Lemmetty, by using an example has illustrated that timing at the sentence level or the correct grouping of words into phrases is difficult because prosodic phrasing is not always marked in text by punctuation, and phrasal accentuation is almost never marked. For example, the input string *John says Peter is a liar* can be spoken in two different ways, giving two different meanings, as in "John says, 'Peter is a liar'" or "'John,' says Peter, 'is a liar.'" In the first sentence Peter is a liar, while in the second the liar is John (Lemmetty, 1999; Möbius, 1999; Möbius et al., 1997). Schweitzer and Möbius (2004) describe some results of exemplar-based production of prosody.

SOME SPEECH SYNTHESIS SYSTEMS

- **Bell Lab TTS:** Bell Labs (now a part of Lucent Technologies) has been active in speech synthesis research for 7 decades and has implemented an interactive and multilingual TTS site on the Web. Bell Labs' systems are concatenative and use diphones as their basic stored speech elements (http://www.bell-labs.com/news/1997/march/5/1.html).
- **DECtalk:** DECtalk, originally developed for telephony applications by the Digital Equipment Corporation's Assistive Technology Group (ATG), uses a digital formant synthesizer. DECtalk supports nine preprogrammed voices: four male, four female, and one child's voice. The choice of a voice, the speech rate, and left- and right-channel stereo audio may be controlled from an Application Programming Interface (API) or in-line text commands (http://research.compaq.com/wrl/DECarchives/DTJ/DTJK01/).
- **ETI-Eloquence:** ETI-Eloquence, based on formant technology, offers TTS conversion in some 13 languages, plus a number of regional variations and dialects, with five built-in voices. Its relatively low memory requirement makes it suitable for use in embedded devices such as mobile phones, in-car navigation systems, and handheld and wireless devices (http://www.scansoft.com/eti/).
- **Eurovocs:** Eurovocs is a stand-alone text-to-speech synthesizer that can be connected to a computer via a standard serial interface (RS232). Each Eurovocs device can support two different languages (http://www.speech.cs.cmu.edu/comp.speech/Section5/Synth/eurovocs.html).
- **Festival:** Festival is a public-domain, multilingual speech synthesis system developed at the Centre for Speech Technology Research (CSTR) at the University of Edinburgh. It offers an environment for research into and development of speech synthesis techniques (http://www.cstr.ed.ac.uk/projects/festival/).
- **IBM TTS:** WebSphere® Voice Server 2.0, an IBM product, offers three alternative TTS technologies, namely, a formant-based voice synthesizer, a concatenative system, and a phrase-splicing system, which provides a range of prerecorded phrases using human voices (http://www.infofax.com/infofax-products/websphere.html).
- **MBROLA:** MBROLA, developed at TCTS (The Circuit Theory and Signal Processing Lab) in the Faculté Polytechnique de Mons in Belgium, is a concatenative TTS system with diphone databases for American, British, and Breton English, Brazilian Portuguese, Dutch, French, German, Romanian, and Spanish in both male and female voices. Several other languages are under development. The MBROLA project aims to develop multilingual speech synthesis for noncommercial purposes and stimulate academic research (http://tcts.fpms.ac.be/synthesis/mbrola.html).
- **PlainTalk:** PlainTalk, available on Apple platforms, uses concatenative synthesis and offers

three adjustable quality levels for slower machines.

- **SVOX:** SVOX TTS is the commercial implementation of a text-to-speech system developed at the Swiss Federal Institute of Technology in Zurich (ETH Zürich). The SVOX TTS engine is designed to be language independent and can be integrated into applications requiring TTS capability. It is multilingual capable using the same voice and even allows for multiple languages within the same sentence (http://www.svox.com).
- **Microsoft Speech Server 2004:** Microsoft Speech Server includes both speech-recognition and TTS capabilities. It uses the Scansoft Speechify TTS engine (http://www.microsoft.com/speech/).
- **Microsoft Reader:** Microsoft Reader incorporates a TTS engine for reading e-books. It can be downloaded from http://www.microsoft.com/reader/downloads/default.asp.

QUALITY OF SYNTHETIC SPEECH

The quality of synthetic speech is measured based on its intelligibility, accuracy, and naturalness. Evaluation procedures may include subjective listening tests with a response set of syllables, words, and sentences, or with a list of questions. The test material is usually focused on consonants because their synthesis is more problematic than that of vowels (Lemmetty, 1999). The quality of a TTS system also depends on how it handles foreign words, acronyms, abbreviations, addresses, homographs, punctuation, and exceptions in the pronunciation rules for numbers.

- **Naturalness:** Naturalness (the way humans speak in conversation) is a primary evaluation criterion for TTS products. With formant synthesis technology, where the speech is generated based on a model of the human voice, naturalness has varied considerably from one product to the other. In the late 1990s, RealSpeak, one of the early concatenative speech synthesis products introduced by L&H (now ScanSoft), appeared in the market. This was a distinct breakthrough and very closely approached human speech in terms of its naturalness. Aculab, AT&T, Babel Technologies, Cepstral, Elan Speech, Fonix, IBM, Microsoft, Nuance, Loquendo, Rhetorical Systems, SpeechWorks, SVOX, and Voiceware subsequently introduced other TTS products based on concatenative technology (http://www.asrnews.com/ttsap/ttspap11.htm).
- **Accuracy:** The primary goal of TTS is to communicate information accurately to the intended users. Apart from pronunciation, the basic letter-to-sound rules and the handling of numbers, acronyms, abbreviations, words of foreign origin, and names are factors in accuracy. The accuracy of TTS products remained relatively static during the period in which their naturalness was improving so dramatically (http://www.asrnews.com/ttsap/ttspap11.htm).

CONCLUSION

Text-to-speech synthesis has developed steadily over the past several decades and has become an important tool. In this article, the techniques used in the implementation of TTS systems and the challenges involved in developing a natural-sounding system have been discussed. Research in this field is ongoing, and we can expect steady improvements in the quality of TTS systems and the appearance of an ever-increasing range of applications over the coming years.

REFERENCES

Allen, J., Hunnicutt, S., Klatt, D.Armstrong , R., & Pisoni, D. (1987). *From text to speech: The MITalk system*. Cambridge University Press Inc, New York, NY, USA.

Black, A., & Lenzo, K. (2003). Building synthetic voices. *Language Technologies Institute, Carnegie Mellon University and Cepstral, LLC.* Retrived May 2004, from *http://www.festvox.org/festvox/bsv-intro-ch.html#AEN61*

Black, A. W., Clark, R., King, S., & Richmond, K. (2004). *The Festival Speech Synthesis System.* Retrieved from *http://www.cstr.ed.ac.uk/projects/festival/*

Carvalho, P., Trancoso, I., & Oliveira, L. (1998). Automatic segment alignment for concatenative speech synthesis in Portuguese. *Tenth Portuguese Conference on Pattern Recognition*, RECPAD98, IST/UTL, Lisbon, Portugal.

Donovan, R. (1996). *Trainable speech synthesis*. PhD thesis, Engineering Department, Cambridge University, Cambridge, United Kingdom. Retrieved March 2004, from *ftp://svr-ftp.eng.cam.ac.uk/pub/reports/donovan_thesis.ps.Z*

Dutoit, T. (1997). *An introduction to text-to-speech synthesis.* Dordrecht: Kluwer Academic Publishers.

Kishore, S. P., Black, A. W., Kumar, R., & Sangal, R. (2003). *Experiments with unit selection speech databases for Indian languages*. National Seminar on Language. Technology Tools: Implementation of Telugu, Hyderabad, India.

Kishore, S. P., Kumar, R., & Sangal, R. (2002). A data-driven synthesis approach for Indian languages using syllable as basic unit. *International Conference on Natural Language Processing (ICON) 2002*, Mumbai, India.

Kominek, J., Bennett, C., & Black, A. (2003). Evaluating and correcting phoneme segmentation for unit selection synthesis. *Eurospeech 2003*, Geneva, Switzerland.

Lemmetty, S. (1999). *Review of speech synthesis technology.* Master's thesis, Helsinki University of Technology, Helsinki, Finland. Retrieved January 2004, from *http://www.acoustics.hut.fi/~slemmett/dippa/*

Möbius, B. (1999). The bell labs German text-to-speech system. *Computer Speech and Language,*13, 319-358.

Möbius, B., Sproat, R., van Santen, J., & Olive, J. (1997). The Bell Labs German text-to-speech system: An overview. *Proceedings of the European Conference on Speech Communication and Technology, 5*, 2443- 2446.

Mukherjee, N., Rajput, N., Subramaniam, L. V., & Verma, A. (2000). On deriving a phoneme model for a new language. *Sixth IEEE International Conference on Spoken Language Processing*, Beijing, China.

Schafer, R. W., & Markel, J. D. (Eds.). (1979). *Speech analysis.* New York: IEEE Press.

Schweitzer, A., Braunschweiler, N., Klankert, T., Möbius, B., & Säuberlich, B. (2003). Restricted unlimited domain synthesis. *Proceedings of Eurospeech 2003*, 1321-1324.

Schweitzer, A., & Möbius, B. (2004). Exemplar-based production of prosody: Evidence from segment and syllable durations. *Speech Prosody 2004*, 459-462.

Stork, D. G. (Ed.). (1996). *The talking computer.* Retrieved February 2004, from *http://mitpress.mit.edu/e-books/Hal/chap6/six1.html*

Syed, M. R., Chakrobartty, S., & Bignall, R. J. (2004). A framework for a Bangla concatenative text-to-speech synthesis system. *Proceedings of IRMA International Conference* (May 2004), New Orleans, LA.

KEY TERMS

Articulatory Synthesis: Articulatory synthesis is the production of speech based on an actual model of the human vocal tract. Such an approach is very complex and computationally intensive. It is almost impossible to model the vocal chords, tongue movements, or other characteristics of the human vocal system perfectly (Lemmetty, 1999).

Concatenative Synthesis: Concatenative synthesis is the combining of a sequence of prerecorded natural utterances from an auditory database to produce arbitrary synthetic speech. In concatenative synthesis, the collecting and labeling of speech samples is very time consuming and may result in a large database. However, the amount of data may be reduced using compression methods. Concatenation points may cause distortion in the speech that is produced (Lemmetty, 1999).

Formant: The formant with the highest energy is called *f*1, the second *f*2, and the third *f*3. Most often

the two first formants, *f*1 and *f*2, are enough to disambiguate a vowel. Sonograms are used to visualise formants (http://www.fact-index.com/f/fo/formant.html).

Formant Synthesis: Formant synthesis entails the use of electronic resonators and filters to emulate the sound produced by the human vocal system. In formant synthesis, a large set of rules is needed to control the formant frequencies and amplitudes, and the excitation source. Some lack of naturalness, especially with nasalized sounds, is considered to be a major problem with formant synthesis (Lemmetty, 1999).

Intonation: Intonation means how the pitch pattern (the property of sound that varies with variation in the frequency of vibration) or fundamental frequency (F0) changes during speech. It refers to the rise and fall of the voice pitch (http://en.wikipedia.org/wiki/Intonation).

Linear Predictive Coding: Linear predictive coding (LPC) is a tool used in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form using a linear predictive model. It is one of the most powerful speech-analysis techniques, one of the most useful methods for encoding good-quality speech at a low bit rate, and it provides extremely accurate estimates of speech parameters (http://www.fact-index.com/l/li/linear_predictive_coding.html).

Phonetics: Phonetics is defined as the study of the articulation of phonemes. It is concerned with the actual nature of the sounds and their production. The meaning of what is articulated is not relevant at this level of linguistic analysis. The objects of study in phonetics are called phones. Phones are actual speech sounds as uttered by humans. Phonetics has three main branches: articulatory phonetics, concerned with the positions and movements of the lips, tongue, and other speech organs in producing speech; acoustic phonetics, concerned with the properties of the sound waves generated; and auditory phonetics, concerned with speech perception (http://www.fact-index.com/p/ph/phonetics.html).

Phonology: Phonology is the study of the sound system of a given language, and the analysis and classification of its abstract linguistic units, called phonemes. A phoneme is the smallest meaningful, contrastive unit in a language and is thus defined on a functional rather than an acoustic or physiological basis. Phonemes have no independent existence: They constitute a structured set in which each element is intentionally different from all of the others (Dutoit, 1997).

Prosody: Intonation, stress, and duration are features of speech that together are called prosodic or suprasegmental features. They determine the melody, rhythm, and emphasis of the speech at the perceptual level. The prosody of continuous speech depends on many separate factors, such as the meaning of the sentence and the speaker's characteristics, intentions, and emotions (Lemmetty, 1999).

2G–4G Networks

Shakil Akhtar
United Arab Emirates University, UAE

INTRODUCTION

The fourth-generation wireless mobile systems, commonly known as 4G, is expected to provide global roaming across different types of wireless and mobile networks; for instance, from satellite to mobile networks and to Wireless Local Area Networks (WLANs). 4G is an all IP-based mobile network using different radio access technologies and providing seamless roaming and connection via always the best available network (Zahariadis & Kazakos, 2003). The vision of 4G wireless/mobile systems will be the provision of broadband access, seamless global roaming and Internet/data/voice everywhere, utilizing for each the most "appropriate" always-best connected technology (Gustafsson & Jonsson, 2003). These systems are about integrating terminals, networks and applications to satisfy increasing user demands (Ibrahim, 2002; Lu & Berezdivin, 2002). 4G systems are expected to offer a speed of more than 100 Mbps in stationary mode and an average of 20 Mbps for mobile stations, reducing the download time of graphics and multimedia components by more than 10 times compared to currently available 2 Mbps on *3G* systems.

Currently, the 4G system is a research-and-development initiative based upon 3G, which is having trouble meeting its performance goals. The challenges for development of 4G systems depend upon the evolution of different underlying technologies, standards and deployment. This article presents an overall vision of the 4G features, framework and integration of mobile communication. First we explain the evolutionary process from *2G* to *4G* in light of used technologies and business demands. Next, we discuss the architectural developments for 2G-4G systems, followed by a discussion on standards and services. Finally, we address the market demands and discuss the development of terminals for these systems.

2G-4G NETWORKS: EVOLUTION

The first generation of *mobile phones* was analog systems that emerged in the early 1980s (Falconer, Adachi & Gudmundson, 1995). The second generation of digital mobile phones appeared in the 1990s, along with the first digital mobile networks. During the second generation, the mobile telecommunications industry experienced exponential growth in terms of both subscribers and value-added services. Second-generation networks allow limited data support in the range of 9.6 kbps to 19.2 kbps. Traditional phone networks are used mainly for voice transmission, and are essentially circuit-switched networks.

2.5G networks, such as General Packet Radio Service (*GPRS*), are an extension of 2G networks in that they use circuit switching for voice and packet switching for data transmission, resulting in its popularity, since packet switching utilizes bandwidth much more efficiently. In this system, each user's packets compete for available bandwidth, and users are billed only for the amount of data transmitted.

3G networks were proposed to eliminate many problems faced by 2G and 2.5G networks, especially the low speeds and incompatible technologies, such as Time Division Multiple Access (*TDMA*) (Falconer, Adachi & Gudmundson, 1995) and Code Division Multiple Access (*CDMA*) (Kohno, Meidan & Milstein, 1995) in different countries. Expectations for 3G included increased bandwidth, 128 Kbps for mobile stations and 2 Mbps for fixed applications (Lu, 2000). In theory, 3G should work over North American as well as European and Asian wireless air interfaces. In reality, the outlook for 3G is not very certain. Part of the problem is that network providers in Europe and North America currently maintain separate standards' bodies (*3GPP* for Europe and Asia; *3GPP2* for North America). The standards' bodies have not resolved the differences in air interface technologies.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

There is also a concern that in many countries 3G will never be deployed due to its cost and poor performance. Although it is possible that some of the weaknesses at the physical layer will still exist in 4G systems, an integration of services at the upper layer is expected.

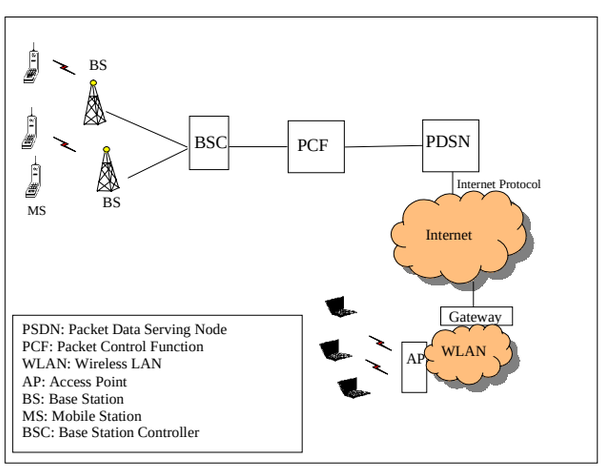
The evolution of mobile networks is strongly influenced by business challenges and the direction mobile system industry takes. It also relates to the radio access spectrum and the control restrictions over it that vary from country to country. However, as major technical advances are being standardized, it becomes more complex for the industry alone to choose a suitable evolutionary path. Many mobile system standards for Wide Area Networks (WANs) already exists, including popular ones such as Universal Mobile Telecommunications Systems (*UMTS*), CDMA and *CDMA-2000* (1X/3X). In addition, there are evolving standards for Personal Area Networks (PANs), such as *Bluetooth* wireless, and for WLANs, such as IEEE 802.11.

The current trend in mobile systems is to support the high bit-rate data services at the downlink via High Speed Downlink Packet Access (HSDPA). It provides a smooth evolutionary path for UMTS networks to higher data rates in the same way as Enhanced Data rates for Global Evolution (*EDGE*) do in Global Systems for Mobile communication (*GSM*). HSDPA uses shared channels that allow different users to access the channel resources in packet domain. It provides an efficient means to share a spectrum that provides support for high data-rate packet transport on the downlink, which is well adapted to urban environment and indoor applications. Initially, the peak data rates of 10 Mbps may be achieved using HSPDA. The next target is to reach 30 Mbps, with the help of antenna array processing technologies, followed by the enhancements in air interface design to allow even higher data rates.

Another recent development is a new framework for mobile networks that is expected to provide multimedia support (Short & Harrison, 2002; Thom, 1996) for IP telecommunication services, called IP Multimedia Subsystems (IMS) (Faccin, Lalwaney & Patil, 2004). Real-time rich multimedia communication mixing telecommunication and data services could happen due to IMS in wireline broadband networks. However, mobile carriers cannot offer their customers the freedom to mix multimedia components (text, pictures, audio, voice, video) within one call. Today, a two-party voice call cannot be extended to a multi-party audio and video conference. IMS overcomes such limitations and makes these scenarios possible.

The future of mobile systems is largely dependent upon the development of 4G systems, multimedia networking and, to some extent, *photonic networks*. It is expected that initially the 4G mobile systems will be developed and used independent from other technologies. With gradual growth of high-speed data support to multimegabits per second, an integration of services will happen. In addition, developments in photonic switching might allow mobile communication on a completely photonic network, using Wavelength Division Multiplexing (*WDM*) on photonic switches and routers.

Figure 1. Wireless mobile system network architecture



Network Architecture

The basic architecture of a wireless mobile system consists of a mobile phone connected to the wired world via single-hop wireless connection to a Base Station (BS), which is responsible for carrying the calls within its region called cell (Figure 1). Due to limited coverage provided by a BS, the mobile hosts change their connecting base stations as they move from one cell to another. A hand-off (later referred to

as "horizontal hand-off") occurs when a mobile system changes its BS. The mobile station communicates via the BS using one of the wireless frequency sharing technologies, such as FDMA, TDMA, CDMA and so forth. Each BS is connected to a Mobile Switching Center (MSC) through fixed links, and each MSC is connected to others via Public Switched Telephone Network (*PSTN*). The MSC is a local switching exchange that handles switching of a mobile user from one BS to another. It also locates the current cell location of a mobile user via Home Location Register (HLR) that stores the current location of each mobile that belongs to the MSC. In addition, the MSC contains a Visitor Locations Register (VLR), with information of visiting mobiles from other cells. The MSC is responsible for determining the current location of a target mobile using HLR, VLR and by communicating with other MSCs. The source MSC initiates a call setup message to an MSC covering a target area for this purpose.

The first-generation cellular implementation consisted of analog systems in 450-900 MHz frequency range using frequency shift keying for signaling and Frequency Division Multiple Access (FDMA) for spectrum sharing. The second-generation implementations consist of TDMA/CDMA implementations with 900 and 1800 MHz frequencies. These systems are called GSM for Europe and IS-136 for US. The respective 2.5G implementations are called GPRS and *CDPD*, followed by 3G implementations.

Third-generation mobile systems are intended to provide a global mobility with a wide range of services including voice calls, paging, messaging, Internet and broadband data. IMT-2000 defines the standard applicable for North America. In Europe, the equivalent UMTS standardization is in progress. In 1998, a Third Generation Partnership Project (3GPP) was formed to unify and continue the technical specification work. Later, the Third Generation Partnership Project 2 (3GPP2) was formed for technical development of CDMA-2000 technology.

3G mobile offers access to broadband multimedia services, which is expected to become all IP based in future 4G systems (Sun Wireless; All IP Wireless – All the Way). However, current 3G networks are not based on IP; rather, they are an evolution from existing 2G networks. Work is going on to provide 3G support and Quality of Service (QoS) in IP and mobility protocols. The situation gets more complex when we consider the WLAN research and expect it to become mobile. It is expected that WLANs will be installed in trains, trucks and buildings. In addition, it may just be formed on an ad-hoc basis (like *ad-hoc networks*; see Chiussi, Khotimsky & Krishnan, 2002; Ramanathan & Redi, 2002; Tseng, Shen & Chen, 2003) between random collections of devices that happen to come within radio range of one another (Figure 2).

In general, 4G architecture will include three basic areas of connectivity (Honkasalo, Pehkonen, Niemi & Leino, 2002; Maniatis, Nikolouzou & Venieris, 2002; Schrick, 2002; Webb, 2001); PANs (such as Bluetooth), WANs (such as IEEE 802.11) and cellular connectivity. Under this umbrella, 4G will

Figure 2. Mobile system/WLAN integration

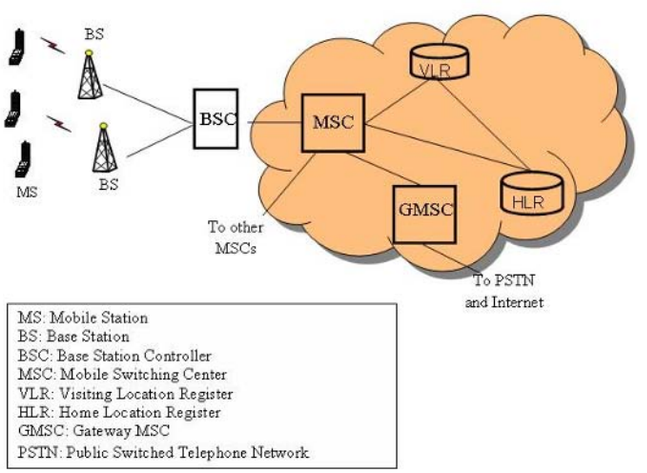
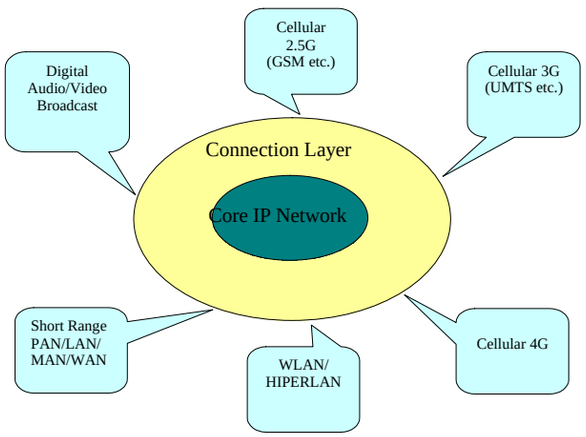


Figure 3: Seamless connection of networks in 4G



provide a wide range of mobile devices that support global roaming (Berezdivin, Breinig & Topp, 2002; Rappaport, Annamalai, Buehrer & Tranter, 2002; Zahariadis, Vaxevanakis, Tsantilas, Zervos & Nikolaou, 2002; Zeng, Annamalai & Bhargava, 1999). Each device will be able to interact with Internet-based information that will be modified on the fly for the network being used by the device at that moment (Figure 3).

In 4G mobile IP, each cell phone is expected to have a permanent "home" IP address, along with a "care-of" address that represents its actual location. When a computer somewhere on the Internet needs to communicate with the cell phone, it first sends a packet to the phone's home address. A directory server on the home network forwards this to the care-of address via tunnel, as in regular mobile IP. However, the directory server also sends a message to the computer, informing it of the correct care-of address, so future packets can be sent directly. This should enable TCP sessions and HTTP downloads to be maintained as users move between different types of networks. Because of the many addresses and multiple layers of subneting, IPv6 is needed for this type of mobility.

The goal of 4G is to replace the current proliferation of core mobile networks with a single worldwide core network standard, based on IPv6 for control, video, packet data and voice. This will provide uniform video, voice and data services to the mobile host, based entirely on IPv6. The objective is to offer seamless multimedia services to users accessing an all IP-based infrastructure through heterogeneous access technologies. IPv6 is assumed to act as an adhesive for providing global connectivity and mobility among networks.

Most of the wireless companies are looking forward to IPv6, because they will be able to introduce new services. The Japanese government is requiring all of Japan's ISPs to support IPv6 by 2006, when the first 4G launch is expected. Although the United States' (U.S.) upgrade to IPv6 is less advanced, WLAN's advancement may provide a shortcut to 4G.

Standards

The role of standards is to facilitate interconnections between different types of telecommunication networks, provide interoperability over network and terminal interfaces, and enable free movement and trade of equipment. Standard bodies in different countries develop telecommunications standards based on government regulations, business trends and public demands. In addition, international standard organizations provide global standardizations. In the telecommunications area, the International Telecommunications Union (ITU) and International Standards Organization (ISO) have been recognized as major international standards developers. Many popular telecommunications and networking standards are given by other international organizations, such as Institute of Electrical and Electronics Engineers (IEEE) and Internet Engineering Task Force (IETF). Among other organizations, the most well known are the Telecommunications Industry Association (TIA) and American National Standards Institute (ANSI) in the U.S., European Telecommunication Standards Institute (ETSI), China Wireless Telecommunications Standards Group (CWTS), Japan's Association of Radio Industries and Businesses (ARIB) and Telecommunications Technology Committee (TTC), and Korea's Telecommunications Technology Association (TTA).

The ITU began its studies on global personal communications in 1985, resulting in a system referred to as International Mobile Telecommunications for the year 2000 (IMT-2000). Later, ITU Radio Communications Sector (ITU-R) and ITU-Telecommunications (ITU-T) groups were formed for radio

Table 1. Comparison of 1G-4G technologies

Technology/ Features	1G	2G	2.5G	3G	4G
Start/ Deployment	1970/ 1984	1980/ 1991	1985/ 1999	1990/ 2002	**2000/ 2006**
Data Bandwidth	1.9 kbps	14.4 kbps	14.4 kbps	2 Mbps	**200 Mbps**
Standards	AMPS	TDMA, CDMA, GSM	GPRS, EDGE, 1xRTT	WCDMA, CDMA-2000	**Single unified standard**
Technology	Analog cellular technology	Digital cellular technology	Digital cellular technology	Broad bandwidth CDMA, IP technology	**Unified IP and seamless combination of broadband, LAN/WAN/ PAN and WLAN**
Service	Mobile telephony (voice)	Digital voice, short messaging	Higher capacity, packetized data	Integrated high-quality audio, video and data	**Dynamic information access, wearable devices**
Multiplexing	FDMA	TDMA, CDMA	TDMA, CDMA	CDMA	**CDMA**
Switching	Circuit	Circuit	Circuit for access network and air interface; Packet for core network and data	Packet except circuit for air interface	**All packet**
Core Network	PSTN	PSTN	PSTN and Packet network	Packet network	**Internet**
Handoff	Horizontal	Horizontal	Horizontal	Horizontal	**Horizontal and Vertical**

communications and telecommunications standards, respectively. In Europe, the concepts of Universal Mobile Telecommunications System (UMTS) have been the subject of extensive research. In 1990, ETSI established an ad-hoc group for UMTS that focused on the critical points to be studied for systems suitable for mobile users. Since 1998, ETSI's standardization of the 3G mobile system has been carried out in the 3G Partnership Project (3GPP) that focuses on the GSM-UMTS migration path. The 3GPP2 is an effort headed by ANSI for evolved IS-41 networks and related radio transmission technologies.

The standard organizations propose the mobile system standards that change as new technologies emerge and the regulations and market demand change. The changing features and used technologies from first- to fourth-generation mobile systems are summarized in Table 1. It is noticeable that the fourth-generation system not only provides a horizontal handoff like the previous systems but also a vertical handoff. While a global roaming may be provided by satellite systems, a regional roaming by 4G cellular systems, local area roaming by WLANs and personal area roaming by wireless PANs, it will also be possible to roam vertically between these systems.

One technology (or its variation) expected to remain in future mobile systems is CDMA, which is a use of the *spread spectrum* technique by multiple transmitters to send signals simultaneously on the same frequency without interference to the same receiver. Other widely used multiple access techniques are TDMA and FDMA, mostly associated with 3G and previous systems.

In these three schemes (CDMA, TDMA, FDMA), receivers discriminate among various signals by the use of different codes, time slots and frequency channels, respectively. Digital cellular systems are an extension of the IS-95 standard and the first CDMA-based digital cellular standard pioneered by Qualcomm. The brand name for IS-95 is cdmaOne. It is now being replaced by IS-2000 and is also known as CDMA-2000, which is a 3G mobile telecommunications standard from ITU's IMT-2000. CDMA-2000 is considered an incompatible competitor of the other major 3G standard, *WCDMA*.

Due to its importance in future systems, let's now examine the different CDMA standards currently available. CDMA-2000 1x, also known as CDMA-2000 1xMC (Multi-Carrier), is the core 3G CDMA-2000 technology. The designation Multi-Carrier refers to the possibility of using up to three separate 1.25 MHz carriers for data transmission, and is used to distinguish this from WCDMA.

WCDMA is the wideband implementation of the CDMA multiplexing scheme, which is a 3G mobile communications standard tied with the GSM standard. WCDMA is the technology behind UMTS.

CDMA-2000 1xRTT (Radio Transmission Technology) is the basic layer of CDMA-2000, which supports up to 144 Kbps packet data speeds. 1xRTT is considered mostly 2.5G technology, which is used to describe systems that provide faster services than 2G, but not quite as fast or advanced as newer 3G systems. CDMA-2000 1xEV (Evolution) is CDMA-2000 1x with High Data Rate (HDR) capability added. 1xEV is commonly separated into two phases, CDMA-2000 1xEV-DO and CDMA-2000 1xEV-DV. CDMA2000 1xEV-DO (Evolution-Data Only) supports data rates up to 2.4 Mbps. It is generally deployed separately from voice networks in its own spectrum. CDMA2000 1xEV-DV (Evolution-Data and Voice) supports circuit and packet data rates up to 5 Mbps. It fully integrates with 1xRTT voice networks. CDMA2000 3x uses three separate 1.25 MHz carriers. This provides three times the capacity but also requires three times more bandwidth.

Network Services

Users relate to different systems with the help of available applications and services that are directly a function of available data rates. The key difference between the 2G and 3G is the data-rate support enabling the later to provide interactive video communication, among other services. A type of service that gained popularity in 2G systems is the messaging service known as Short Messaging Services (SMS), which is a text messaging service for 2G and later mobile phones. The messages in SMS cannot be longer than about 160 characters. An enhanced version of SMS, known as Enhanced Messaging Service (EMS), supports the ability to send pictures, sounds and animations. A newer type of messaging service, Multimedia Messaging Service (MMS), is likely to be very popular for 3G systems and beyond. MMS provides its users the ability to send and receive messages consisting of multimedia elements from person to person as well as the Internet, and serves as the e-mail client. MMS uses Wireless Application Protocol (*WAP*) technology and is powered by the well-known technologies EDGE, GPRS and UTMS (using WCDMA). The messages may include any combination of text, graphics, photographic images, speech and music clips or video clips.

The most exciting extension of messaging services in MMS is a video message capability. For instance, a short 30-second video clip may be shot at a location, edited, with appropriate audio added, and transmitted with ease using the mobile keys on the cellular phones. In addition, by using Synchronized Multimedia Integration Language (SMIL), small presentations can be made that incorporate audio and video along with still images, animations and text to assemble full multimedia presentation by using a media editor.

With MMS, a new type of service, Interfacing Multimedia Messaging Services (IMMS), is expected to emerge that integrates MMS and Mobile Instant Messaging (MIM), allowing users to send messages in their MIM buddy list. This will bring full integration of state-of-the art mobile messaging services including MIM, MMS and chat into all types of mobile devices.

A new term, “Mobile Decision Support” (MDS) has been coined recently for a unique set of services and applications that will provide instant access to information in support of real-time business and personal activities for vehicle-based 3G systems. Some example services are navigation, emergency services, remote monitoring, business finder, e-mail and voice mail. It is expected that MDS-based services will generate a huge non-voice traffic over the Internet.

WAP is an open international standard for applications that use wireless communication on mobile phones. The primary language of the WAP specification is Wireless Markup Language (WML), which is the primary content based on XML (a general-purpose markup language to encode text, including the details about its structure and appearance). The original intent in WAP was to provide mobile replacement of the World Wide Web. However, due to performance limitations and costs, it did not become quite popular as originally expected.

Although WAP never became popular, a popular WAP-like service called i-mode has recently been developed in Japan that allows Web browsing and several other well-designed services for mobile phones. i-mode is based upon Compact HTML (C-HTML) as an alternate to WML, and is compatible with HTML, allowing the C-HTML Web sites to be viewed and edited using standard Web browsers and tools.

Terminals

A mobile phone system is used as a basic terminal for communication. Also called a wireless phone, handset, cellular mobile or cell phone, it is a mobile communications system that uses a combination of radio-wave transmission and conventional telephone switching to permit telephone communication to and from mobile users within a specified area. A 2.5G/3G terminal may consist of a mobile phone, computer/ laptop, television, pager, videoconferencing center, newspaper, diary or even credit card. Often, these terminals may require a compatible 3G card and specialized hardware to provide the desired functionality.

The terminal design considerations are influenced by the potential applications and bandwidth requirements. However, there are standards for mobile stations specifications as well; for instance, in IMT-2000. The actual mobile design varies based mainly upon the multiple available standards, speeds, displays and operating systems. Numerous smartphone operating systems are tailored to specific products by well-known companies such as Palm, Nokia, Sony, Ericsson, Siemens, Alcatel, Motorola, Samsung, Sanyo, Panasonic, Mitsubishi, LG, Sharp, Casio, NEC, NTT DoCoMo, KDDI and so forth. The key capability in most of the supported products is the camera, video clips, keyboards, touchscreen, voice recognition, WiFi (IEEE 802.11b WLAN) and bluetooth wireless.

Future trends in wireless terminals include the influences of new technology such as software radio, wireless socket (WiFi), portability and new design/ display concepts. The newer smartphones are expected to have the functionalities of a Pocket PC with features such as Pocket Outlook, Pocket Internet Explorer, Windows Media Player and MSN Messenger. These newer services will obviously make the communication in 4G systems much easier. However, the biggest challenge remains the integration and convergence of the technologies at the lower layers.

CONCLUSION

This article considers the current and future trends in mobile systems, including the evolutionary path – starting from 1G mobile phone systems and continuing to the development of 4G systems. Evolution in network design, architecture, standards, services and terminals are discussed, as well.

REFERENCES

Berezdivin, R., Breinig, R., & Topp, R. (2002). Next generation wireless communications concepts and technologies. *IEEE Communications Magazine*, *40*(3), March, 108-116.

Chiussi, F.M., Khotimsky, D.A., & Krishnan, S. (2002). Mobility management in third generation all-IP networks. *IEEE Communications Magazine, 40*(9), September, 124-135..

Faccin, S.M., Lalwaney, P., & Patil, B. (2004, January). IP multimedia services: Analysis of mobile IP and SIP interactions in 3G networks. *IEEE Communications Magazine*, *42*(1), 113-120.

Falconer, D., Adachi, F., & Gudmundson, B. (1995, January). Time division multiple access methods for wireless personal communications. *IEEE Communications Magazine*, 33(1), 50-57.

Gustafsson, E., & Jonsson, A. (2003). Always best connected. *IEEE Wireless Communications*, February, 49-55.

Honkasalo, H., Pehkonen, K., Niemi, M.T., & Leino, A.T. (2002). WCDMA and WLAN for 3G and beyond. *IEEE Wireless Communications*, *9*(2), April, 14-18.

Ibrahim, J. (2002). 4G features. *Bechtel Telecommunications Technical Journal*, *1*(1), December, 11-14.

Kohno, R., Meidan, R., & Milstein, L.B. (1995). Spread spectrum access methods for wireless communications. *IEEE Communications Magazine*, January.

Lu, W.W. (Guest Ed.) (2000). Third generation wireless mobile communications & beyond. *IEEE Personal Communications*, *7*(6), December, 5-47.

Lu, W.W., & Berezdivin, R. (Guest Eds.) (2002). Technologies on fourth generation mobile communications. *IEEE Wireless Communications*, *9*(2), April, 8-71.

Maniatis, S.I., Nikolouzou, E.G., & Venieris, I.S. (2002, August). QoS issues in the converged 3G wireless and wired networks. *IEEE Communication Magazine, 40*(8), 44-53.

Ramanathan, R., & Redi, J. (2002). A brief overview of ad-hoc networks: Challenges and directions. *IEEE Communications Magazine*, May (50th anniversary issue).

Rappaport, T.S., Annamalai, A., Buehrer, R.M., & Tranter, W.H. (2002). Wireless communications: Past events and a future perspective. *IEEE Communications Magazine*, May (50th anniversary issue).

Schrick, B., &Riezenman, M.J. (2002). Wireless broadband in a box. *IEEE Spectrum*, June.

Short, M., & Harrison, F. (2002). A wireless architecture for a multimedia world. *The Journal of Communications Network*, *1*(1), April-June.

Sun Wireless. All IP wireless, all the time – Building a 4th generation wireless network with open Systems solutions. Retrieved March 1, 2005 from *http://research.sun.com/features/4g_wireless/*

Thom, G.A. (1996). H.323: The multimedia communications standard for Local Area Networks. *IEEE Communications Magazine*, *34*(12), December, 52-56.

Tseng, Y., Shen, C., & Chen, W. (2003). Integrating mobile IP with ad hoc networks. *IEEE Computer Magazine, 36*(5), May, 48-55.

Webb, W. (Ed.) (2001). *The future of wireless communications.* Artech House.

Zahariadis, T. & Kazakos, D. (2003). (R)Evolution toward 4G mobile communication systems. *IEEE Wireless Communications*, *10*(4), August.

Zahariadis, T.B., Vaxevanakis, K.G., Tsantilas, C.P., Zervos, N.A., & Nikolaou, N.A. (2002). Global roaming in next-generation networks. *IEEE Communications Magazine*, *40*(2), February, 145-151.

Zeng, M., Annamalai, A., & Bhargava, V.K. (1999). Recent advances in cellular wireless communications. *IEEE Communications Magazine*, *37*(9), September, 128-138.

WEB SITES

1. Third Generation Partnership Project (3GPP), *www.3gpp.org*
2. Third Generation Partnership Project 2 (3GPP2), *www.3gpp2.org*
3. Wireless Application Protocol (WAP) Forum, *www.wapforum.org*
4. Global Systems for Mobile Communication (GSM) Association, *www.gsmworld.com*
5. European Telecommunications Standards Institute (ETSI), *www.etsi.org*
6. International Telecommunications Union (ITU), *www.itu.org*
7. Code Division Multiple Access (CDMA) Development Group, *www.cdg.org*
8. Internet Engineering Taskforce (IETF), *www.ietf.org*
9. Institute of Electrical and Electronics Engineers (IEEE), *www.ieee.org*
10. American National Standards Institute (ANSI), *www.ansi.org*
11. Telecommunication Standards Institute (TIA), *www.tiaonline.org*
12. Association of Radio Industries and Businesses (ARIB), *www.arib.or.jp*
13. China Wireless Telecommunication Standards (CWTS) group, *www.cwts.org*
14. International Standards Organization (ISO), *www.iso.org*
15. Telecommunications Technology Association (TTA), *www.tta.or.kr*

KEY TERMS

1G: Old-fashioned analog mobile phone systems capable of handling very limited or no data at all.

2G: Second-generation voice-centric mobile phones and services with limited data rates ranging from 9.6 kbps to 19.2 kbps.

2.5G: Interim hardware and software mobile solutions between 2G and 3G, with voice and data capabilities and data rates ranging from 56 kbps to 170 kbps.

3G: Long-awaited digital mobile systems with a maximum data rate of 2 Mbps under stationary conditions and 384 kbps under mobile conditions. This technology is capable of handling streaming video, two-way voiceover IP and Internet connectivity, with support for high-quality graphics.

3GPP: Third Generation Partnership Project. 3GPP is an industry body set up to develop a 3G standard based upon wideband CDMA (WCDMA).

3GPP2: Third Generation Partnership Project 2. 3GPP2 is an industry standard set up to develop a 3G standard based upon CDMA-2000.

3.5G: Interim systems between 3G and 4G allowing a downlink data rate up to 14 Mbps. Also called High Speed Downlink Packet Access (HSDPA).

4G: Planned evolution of 3G technology expected to provide support for data rates up to 100 Mbps, allowing high-quality and smooth-video transmission.

Ad-Hoc Networks: It is a self-configuring mobile network of routers (and hosts) connected by wireless, in which the nodes may move freely and randomly, resulting in a rapid and unpredictable change in network's wireless topology. Also called Mobile Ad-hoc NETwork (MANET).

Bluetooth: It is a wireless networking protocol designed to replace cable network technology for devices within 30 feet. Like IEEE 802.11b, Bluetooth also operates in unlicensed 2.4GHz spectrum, but it only supports data rates up to 1 Mbps.

CDMA: Code Division Multiple Access, also known as CDMA-ONE or IS-95, is a spread spectrum communication technology that allows many users to communicate simultaneously using the same frequency spectrum. Communication between users is differentiated by using a unique code for each user. This method allows more users to share the spectrum at the same time than alternative technologies.

CDMA-2000: Sometimes also known as IS-136 and IMT-CDMA multicarrier (1X/3X), CDMA-2000 is an evolution of narrowband radio transmission technology known as CDMA-ONE (also called CDMA or IS-95), to third generation. 1X refers to the use of 1.25 Mhz channel while 3X refers to 5 Mhz channel.

CDPD: Cellular Digital Packet Data is a wireless standard providing two-way data transmission at 19.2 kbps over existing cellular phone systems.

DSSS: In Direct Sequence Spread Spectrum, the data stream to be transmitted is divided into small pieces, each of which is allocated a frequency channel. Then the data signal is combined with a higher data-rate bit sequence known as a "chipping code" that divides the data according to a spreading ratio, thus allowing resistance from interference during transmission.

EDGE: Enhanced Data rates for Global Evolution technology gives GSM and TDMA the capability to handle 3G mobile phone services with speeds up to 384 kbps. Since it uses the TDMA infrastructure, a smooth transition from TDMA-based systems such as GSM to EDGE is expected.

FHSS: In Frequency Hopping Spread Spectrum, a broad slice of bandwidth spectrum is divided into many possible broadcast frequencies to be used by the transmitted signal.

GPRS: General Packet Radio Service provides data rates up to 115 kbps for wireless Internet and other types of data communications, using packet data services.

GSMC: Global Systems for Mobile Communication is a worldwide standard for digital wireless mobile phone systems. The standard was originated by the European Conference of Postal and Telecommunications Administrations (CEPT), who was responsible for the creation of ETSI. Currently, ETSI is responsible for the development of the GSM standard.

Mobile Phones: Mobile communication systems that use radio communication and conventional telephone switching to allow communication to and from mobile users.

Photonic Networks: A network of computers made up using photonic devices based on optics. The devices include photonic switches, gateways and routers.

PSTN: Public Switched Telephone Network is a regular voice telephone network.

Spread Spectrum: It is a form of wireless communication in which the frequency of the transmitted

signal is deliberately varied over a wide range. This results in a higher bandwidth of the signal than one without varied frequency.

TDMA: Time Division Multiple Access is a technology for sharing a medium by several users by dividing into different time slots transmitting at the same frequency.

UMTS: Universal Mobile Telecommunications System is the 3G mobile telephone standard in Europe proposed by ETSI.

WAP: Wireless Application Protocol defines the use of TCP/IP and Web browsing for mobile systems.

WCDMA: Wideband CDMA is a technology for wideband digital radio communications of multimedia and other capacity-demanding applications. It is adopted by ITU under the name IMT-2000 direct spread.

WDM: Wavelength Division Multiplexing allows many independent signals to be transmitted simultaneously on one fiber, with each signal located at a different wavelength. Routing and detection of these signals require devices that are wavelength selective, allowing for the transmission, recovery or routing of specific wavelengths in photonic networks.

Type Justified

Anna Szabados
Mission College, USA

Nishikant Sonwalkar
Massachusetts Institute of Technology, USA

INTRODUCTION

There is growing global need for quality online education with increased classroom engagement and student-focused teaching approaches. The present text-heavy approach dominating online education is wholly unsatisfactory as a learning experience.

Our Web-based teaching module provides an augmented teaching solution that serves as an online tool for hybrid courses in typography, which is a course of major importance for graphic, multimedia, Web, environmental, and packaging design students.

Our examination of the literature is focused on e-learning, constructivist teaching, and the effective incorporation of multimedia into Web-based teaching/learning.

Establishing the need for a new paradigm of e-learning, we then describe the adaptive learning that enables individualized instruction, based on learning style preferences. The new pedagogical framework providing numerous learning styles, continuous assessment, and remediation leads to an extremely powerful teaching/learning environment for creative art of typography.

BACKGROUND

Technology increasingly is becoming part of mainstream education. Terms such as e-learning, online learning, and Web-based teaching refers to the dissemination of learning content over the Internet with a desired instructional design. These learning modules are then accessed by learners to purse educational activities using Internet browsers and servers.

Wedemayer (1981) identifies the importance of independent learning through a student learning activity. He is a proponent of greater learner freedom.

Lee and Ovens (2000) state that in order to create successful Web-based learning experiences, one must consider the creative skill of the developers, the bandwidth, and the hardware and software capabilities.

Pittinsky (2003) sees great potential for e-learning for several reasons. First, there is a renewed strong focus on the learners' needs, the availability of technological resources, the search for new funding sources, and the opportunity to provide education to new markets.

According to Bates et al. (2003), learning from a computer will have a different effect than learning from books or traditional lectures. Each of these activities provides a different form of knowledge. Deep learning (integrated understanding of concepts) takes place when a learner is able to integrate and reconcile all types of learning. Bates adds that it often is helpful to learn about the same thing in different ways.

Aggarval (2000) states that high-quality design of Web-based courses will add significant educational value. In order to design effectively, one must consider the goals, needs, and characteristics of the target audience. Successful Web teaching/learning experiences include a high level of interaction between instructors and students and also among participating students. The integrity of the educational process depends on two-way communication.

Schank (2002) asserts that if learning is not engaging to the student, no real learning takes place. He states that "doing" is interesting. Doing promotes engagement. Therefore, Web-based e-learning that emphasizes interaction and includes simulations and multimedia provides an optimal learning environment.

According to Mishra and Sharma (2004), the probability of student learning is much higher when

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

students are able to discuss, write, and, most importantly, apply knowledge that is relevant to their daily lives. The project must be complex with no clear-cut answers.

Clark and Mayer (2003) recommend that e-learning courses include both graphics and words, rather than text alone.

They use the term *graphics* to describe a variety of visual elements such as technical illustrations, drawings, charts, diagrams, photos, animations, and video. The use of relevant visuals will foster active learning and increase understanding of the course content. Research shows that images help learners to make mental connections more effectively and provide a deeper learning experience. One of the key terms is *relevant visuals*; in other words, the authors recommend not using images to decorate, but to enlighten and explain information.

A series of tests was done (Mayer, Heiser & Lonn, 2001) using several student groups. Some received information in animation form with concurrent audio; others received the same information with concurrent text. The non-redundant group produced 43% to 69% higher scores on a problem-solving transfer test. Based on these studies, one could make the case that less is more.

In fact, the principle of less is more has been scientifically proven. In design classes, we talk about the Gesthalt principles. The Gestalt school of psychology, which began in Germany around 1912, investigated how we see and organize visual information into a meaningful whole. The conviction developed that the whole cannot be perceived by a simple addition of isolated parts. Each part is influenced by those around it. When we see things that are similar, we naturally group them. Grouping by similarity occurs when we see similar shapes, sizes, colors, spatial location angles or values. In a group of similar shapes or angles, we will notice the dissimilar. This is called the principle of visual anomaly.

The Gesthalt studies also pointed out the fact that, as human beings, we look for patterns and logical connections between visual elements; our understanding increases if we have fewer visual elements to understand. This is quite important since the primary role of design is effective communication, so the principles of effective graphic/multimedia design communication must be transferred in order to produce more effective e-learning.

TYPOGRAPHY MODULE DEVELOPMENT

Typography, the historical examination, design, development, and appropriate use of type fonts, is a subject of great importance for design students specializing in graphic design, Web design, multimedia design, packaging design, and so forth. Type Justified is an interactive, Web-based teaching module that was developed with several goals in mind. The first goal was to provide factual typographic information in an interactive manner and include a built-in evaluation system. The second goal was to create an actual project for graphic and multimedia design students that incorporates effective team-based learning. The third goal was to establish collaboration with an industry partner who could provide evaluation software support.

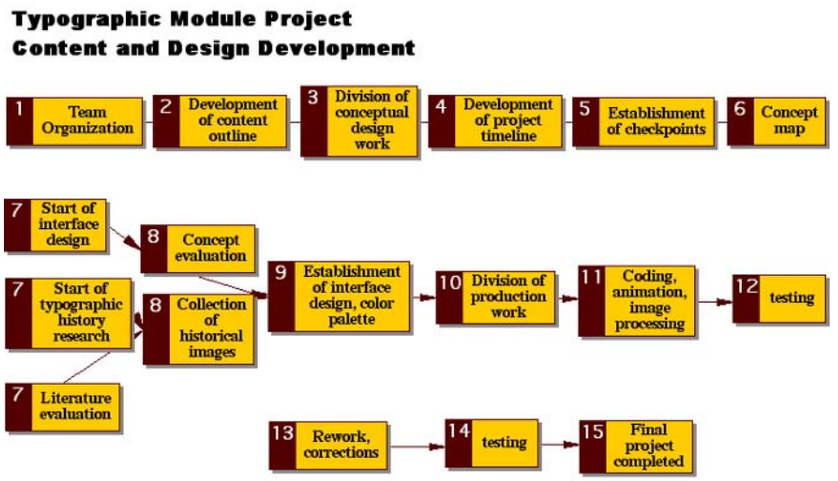
The resulting module achieves these goals and serves to support instructors and students nationwide within graphic design, Web design, multimedia design, packaging design, industrial design, and so forth.

The module was developed using key technologies to help enhance the learning experience, including HTML, Javascript, Macromedia Flash technologies, and the iDesign evaluation software program.

The students had to consider basic Web variables such as the importance of dealing with screen resolution limitations, font selection for the screen, understanding, flexible page size, browser issues, creating graphic or image-based type for the Web, and motion on the computer screen.

Figure 1 identifies the major steps of the module design and development. The six-member advanced design student team was divided into three subgroups that focused on separate aspects of the project. All six team members met twice a week for progress checks, and once a week they met with their instructor. The students and the instructor kept in constant daily contact via e-mail. The content outline and project timeline was developed along with a concept map that clearly identified the necessary tasks. Subgroup 1 started work on the interface design, while subgroups 2 and 3 worked on typographic history research and literature evaluation. The final interface design concept was established along with the color palette. Production work was divided among the subgroups, followed by weekly progress reports and presentations. Web module production included image com-

Figure 1. The step-by-step process of module development


Typographic Module Project
Content and Design Development
1 Team Organization
2 Development of content outline
3 Division of conceptual design work
4 Development of project timeline
5 Establishment of checkpoints
6 Concept map
7 Start of interface design
8 Concept evaluation
7 Start of typographic history research
8 Collection of historical images
9 Establishment of interface design, color palette
10 Division of production work
11 Coding, animation, image processing
12 testing
7 Literature evaluation
13 Rework, corrections
14 testing
15 Final project completed

pression, HTML coding (Hyper Text Markup Language provides tags to mark the texts and graphics objects so Web-browsers can render these elements with correct format on the Web using the Internet), Macromedia Flash animation development, text editing, and so forth. Finally, all the elements were linked, and the module went through another cycle of testing and corrections. Testing took place on multiple platforms and browsers.

The historical content of the module covers the time period from the beginning of Western writing—Sumerian cuneiform to the present. The module also allows students to study typographic terminology in a playful, interactive manner along with the examination of typographic effects, such as the psychological and cultural impacts of different typefaces. The module dedicates separate segments to the study of readability, typographic meaning, and the establishment of visual hierarchy along with examples of using type as a symbol and type used as illustration to clarify content.

Examples of student engagement included collaborative work, ongoing communication among project participants, collection, review and exploration of module content, and incorporation of the design process into the project development. In addition, the project was viewed by the students as an authentic real-world learning experience with a product outcome that has an effect beyond the classroom. In addition, this semester-long project provided a safe, student-centered learning experience.

The project takes into consideration the different learning styles. The incorporation of audio, text, and visuals gives appropriate options to learners. After the completion of the module design, the project was forwarded to the industry partner, IDL Systems, to incorporate the adaptive learning features using the iDesigner software package.

Students who used the module were given several options to study the content, taking into consideration their individual learning preferences, from basic text-based learning to full multimedia discovery-based learning. The built-in evaluation system periodically tests the students' understanding of the subject. If students score low, the system will direct them to the material that needs to be reviewed.

ADAPTIVE LEARNING

The genesis of adaptive learning systems is from the artificial intelligence (AI) research (Sonwalkar). In the early 1980s, there was significant development of systems to provide intelligent response to user interacting with the computers. Adaptive learning systems can be defined as the intelligent systems that are self-organized, based on the observation of the learning preferences of an individual, resulting in the best learning performance.

PROJECT PARTNERS

iDL, the industry partner who worked on this project, is an e-learning technology company that has developed a rapid course creation environment, allowing

Table 2.

Limitations With Existing Online Educational Programs	iDL's Adaptive Learning Solutions
One-size-fits-all approach	Mass Customization technology adapts the course to the needs of each individual learner
Merely a means for displaying course content	Dynamic and interactive content engages and motivates learning
Minimal educational effectiveness	Incorporates strong pedagogical techniques that promote effective learning
Restrictive linear course structure	Five learning styles allow students to learn in their own unique ways
Less than 50% of students pass a typical course	Average passing rate of 95%
Conventional e-learning technology platforms are obsolete	Next Generation technology with adaptive learning and intelligent feedback

dynamic content sequencing to create online courses with multiple learning styles and media, providing a concept-map-based tracking of learners for intelligent feedback and remediation. iDL has developed the next generation of adaptive learning style management systems for providing pedagogically effective asynchronous online learning based on the Learning Cube pedagogical framework developed by Dr. Sonwalkar at MIT (Sonwalkar, 2001, 2002, 2003, 2004).

The technological differentiators for adaptive learning methodology developed by iDL Systems are summarized below:

- Adaptive learning technology that provides online education matching the learning styles of the individual learners (Mass-customization).
- Provide individualized intelligent feedback and remediation courses just-in-time for learning.
- The ease of use and engaging courses have extremely high completion rate (95%+).
- Complete suite of intelligent products that make design, development, and deployment of the online course much better (adaptive), faster (four-week delivery cycle), and cheaper by 50%.
- The courses conform to SCORM 1.2 educational technology standards.

PROJECT PARTICIPANTS

The New Media Center headquartered at Mission College sponsored this project. The typography module was designed and developed with the assistance of six advanced multimedia design students with a faculty facilitator in the award-winning Graphic and Multimedia Design Program at Mission College. The students spent about 600 hours on the project. Industry partner IDL Systems incorporated their iDesigner software package with the work, thereby adding advanced adaptive learning features to the module.

The new paradigm of adaptive learning has shown potential to overcome the difficulties posed by traditional online programs (Sonwalkar, 2004). Most of the existing online courses/programs tend to be text-heavy and do not take advantage of the interactive and/or multimedia capabilities the Web offers, nor do they incorporate different learning styles. Most of the existing online courses also lack effective feedback and evaluation mechanisms.

ADAPTIVE LEARNING SYSTEMS

Adaptive learning systems can be defined as the intelligent systems that are self-organized, based on the observation of the learning preferences of an individual, resulting in the best learning performance.

This definition illustrates the following important characteristics of adaptive learning systems:

1. Adaptive systems need to have a well-defined pedagogical framework in order to identify and differentiate individual learning preferences.
2. The systems need to have a well-defined quantification of learning performance and learning preference inference model.
3. The systems need to have a dynamic content sequencing engine in order to organize learning assets to match the individual learning.

The key advantage of adaptive learning is the high-touch approach provided by the continuous diagnostic tests and immediate adaptive feedback. As a result, online users get a highly interactive environment that is engaging and adaptive, and provides continuous assistance in reaching learning objectives with their own learning styles.

FUTURE TRENDS

The future of e-learning is not in providing static content that just provides information, but lies in the power of customizing the content to match the learning needs of each individual learner. The learning process that is based on strong interactive, visual, engaging presentation and continuous adaptive feed back can overcome the deficiencies prevalent in most current online learning offerings and meet the long-awaited promise of educational revolution.

According to Mishra and Sharma (2004), in the future, multimedia strategies will have a much stronger impact on how instructors engage students in the educational process.

Another growing trend is greater collaboration between industry and educational institutions. According to Lee and Owens (2000), companies that will not invest in the infrastructure for Internet training development and delivery will be left behind and will quickly lose their competitive edge.

Wand (2002) sees the trend on the Web becoming one of increasing commercialization. It will become not only a required promotional tool for corporations, but also a significant part of their business plans and revenues. At the same time, one can look forward to the seamless combinations of text, audio, video and animation, along with the growth of interactive multimedia environments in business, industry, and education.

CONCLUSION

This project is a successful response to several educational goals. The highly visual and interactive module provides factual typographic information and includes a built-in adaptive learning and evaluation system. This project became a real-life project for advanced-level graphic and multimedia design students that incorporated effective team-based learning. The module also served as an example of successful collaboration between an educational institution and an industry partner.

REFERENCES

Aggarwal A. (2000). *Web-based learning and teaching technologies: Opportunities and challenges (Vol. 1)*. Hershey, PA: Idea Group Publishing.

Bates, G.P. (2003). *Effective teaching with technology in higher education (Vol. 1)*. San Farncisco, CA: Jossey Bass.

Mishra, S., & Sharma, R.C. (2004). *Interactive multimedia in education and training (Vol. 1)*. Hershey, PA: Idea Group Publishing.

Pittinsky, M.S. (2003). *The wired tower. Perspectives on the impact of the Internet on higher education (Vol. 1)*. Upper Saddle River, NJ: Prentice Hall.

Ruth, C., & Clark, R.E.M. (2003). *E-learning and the science of instruction (Vol. 1*). San Francisco: Pfeiffer.

Schank, R.C. (2002). *Designing world-class e-learning (Vol. 1)*. New York: McGraw Hill.

Sonwalkar, N. (2001a). The sharp edge of the cube: Pedagogically driven instructional design for online education. *Syllabus, 12*, 12-16.

Sonwalkar, N. (2001b). Optimizing individual learning performance with multi-dimensional evaluations and adaptive systems. *NASA/CP-2003-212437*.

Sonwalkar, N. (2002). A new methodology for evaluation: The pedagogical rating of online courses. *Syllabus, 1*, 18-21.

Sonwalkar, N. (2004a). Adaptive learning: The next generation of online learning. *Proceedings of The Future is Now*, Salem Massachusetts.

Sonwalkar, N. (2004b). *Changing the interface of education with revolutionary learning technologies (Vol. 1)*. New York: iUniverse Publishing Inc.

Wands, B. (2002). *Digital creativity (Vol. 1)*. New York: John Wiley & Sons Inc.

Wedemayer, C. (1981). *Learning at the backdoor*. Madison, WI: University of Wisconsin Press.

KEY TERMS

Compression: It is necessary to reduce file size so an image can download more quickly on the Web.

HTML: Hypertext Markup Language. The most often used coding language used on the World Wide Web. It uses basic word processing tags to specify formatting, linking, and so forth during the creation of Web pages. HTML requires no special Web layout software.

Interface: The design on the computer-screen with which the user interacts.

Macromedia Flash: Vector-based animation software program produced by Macromedia Corporation.

Multimedia: The combination of text, images, sound, and moving images to create an interactive experience.

Static Graphics: Graphics with no animation or interactivity. The computer-image equivalent of a photograph or a painting.

Typeface: The particular style and design of alphabetic letters, numbers, and symbols that make up a font.

Typography: The art of selecting, designing, and using appropriate typefaces in the content of the material in order to produce clear, legible, and aesthetically appealing reading material for print or screen.

Vector Graphic: A vector graphic file contains all the calculations to redraw an image onscreen. A vector graphic's file size remains small, and the image can be scaled to any size without any degradation to image quality.

Web-Based Teaching: All or most teaching takes place on the Web with no or little face-to-face interaction.

Ubiquitous Commerce

Holtjona Galanxhi-Janaqi
University of Nebraska-Lincoln, USA

Fiona Fui-Hoon Nah
University of Nebraska-Lincoln, USA

UBIQUITOUS COMMERCE: THE NEW WAVE

Ubiquitous commerce, also referred to as u-commerce or übercommerce, is the combination of electronic, wireless-mobile, television, voice, and silent commerce. However, its full realization would bring something more than the simple sum of its components. Ubiquitous commerce can be defined as "the use of ubiquitous networks to support personalized and uninterrupted communications and transactions between a firm and its various stakeholders to provide a level of value, above, and beyond traditional commerce" (Watson, Pitt, Berthon, & Zinkhan, 2002).

CHARACTERISTICS OF UBIQUITOUS COMMERCE

One of the characteristics of u-commerce is ubiquity. It means that computers will be everywhere and every device will be connected to the Internet. It is this omnipresence of computer chips that will make them "invisible" as people will no longer notice them (Watson et al., 2002).

U-commerce will also add universality. Universality will eliminate the problems of incompatibility caused by the lack of standardization like the use of mobile phones in different networks. A universal device will make it possible to stay connected at anyplace and anytime.

U-commerce will add uniqueness of information. Uniqueness means that the information provided to the users will be easily customized to their current context and particular needs in a specific time and place.

Finally, unison aggregates the aspects of application and data into one construct (Junglas & Watson, 2003b). In a u-commerce environment, it is possible to integrate various communication systems such that there is a single interface or connection point to them (Watson et al., 2002).

COMPONENTS OF UBIQUITOUS COMMERCE

Junglas and Watson (2003a) view u-commerce as a conceptual extension of e-commerce and m-commerce (mobile commerce).

Electronic Commerce

Electronic commerce is the use of the Internet and the Web to transact business. There are three main types of e-commerce: business to consumer, business to business, and consumer to consumer. In addition, government-to-government, government-to-consumer, and consumer-to-government e-commerce have emerged. E-commerce is the most established type of commerce performed through digital means. Companies are using it as a part of their traditional commerce or as a pure online business model.

Wireless Commerce

Wireless commerce extends e-commerce with characteristics such as reachability, accessibility, localization, identification, and portability. Wireless commerce is a key part of u-commerce because it creates the possibility for communications between people, businesses, and objects to happen anywhere and anytime. Mobile and wireless devices are enabling organizations to conduct business in more efficient and effective ways (Nah, Siau, & Sheng, 2005). Wireless devices can offer many advantages for

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

companies and individuals such as empowering the sales force, coordinating remote employees, giving workers mobility, improving customer service, and capturing new markets.

Other components of u-commerce are voice, television, and silent commerce.

Voice Commerce

An increasing number of businesses are using computerized voice technologies: speech recognition, voice identification, and text to speech. Voice commerce enables businesses to reduce call-center operating costs and improve customer service. Voice commerce can also be used to generate new sources of revenue, but this will probably take longer to materialize. Companies are mostly pursuing voice commerce as a part of a multichannel strategy.

Television Commerce

The spread of interactive digital television will provide a platform for two-way personalized communication in the center of most homes. This will make television commerce a big opportunity for business and a critical component of u-commerce. Television commerce is mainly used as an end-consumer channel. Since it can reach a big range of the population, governments may also use it to deliver their services. Digital television is also a suitable method to deliver innovative services. Interacting TV (TiVo) integrates software and set-top boxes to facilitate digital interactive television with many capabilities, including time-shifting content and filtering advertisements.

Silent Commerce

Silent commerce refers to the business opportunities created by making everyday objects intelligent and interactive. For example, radio-frequency identification (RFID) chips allow the tagging, tracking, and monitoring of objects along an organization's supply chain. An important advantage of RFID as compared to technologies like bar codes is its ability to identify and track individual assets while barcodes can only identify classes of assets. Microelectromechanical systems (MEMSs) chips combine the capabilities of an RFID tag with small, embedded mechanical devices such as sensors. Nowadays, researchers are even talking about nanoelectromechanical systems (NEMSs) or structures, which have dimensions below a micron.

With more advanced silent commerce applications, it will be possible for organizations to identify, track, and monitor every single product along the entire supply chain and even after the sale, up to the point when the product is recycled. These more complex solutions could completely transform the businesses of tomorrow and they create a stream of information and value.

THE DRIVERS FOR THE GROWTH OF U-COMMERCE

Schapp and Cornelius (2001) identify three global phenomena that will accelerate the growth of u-commerce.

Pervasiveness of Technology

The explosive growth of nanotechnology and the continuing capital investments in technology at the enterprise level increase the pervasiveness of the technology and expand the platform on which to leverage innovation and new applications. Two of the main barriers are size and power supply. Bluetooth and MEMS technology are examples that can help overcome these barriers in the future.

Growth of Wireless

Wireless is one of the fastest-growing distributed bases: Wireless networks have expanded around the globe, and mobile phone usage and new applications have exploded. Wireless commerce is therefore a critical component of u-commerce. It is critical to resolve the related issues in order to capture the full advantage that u-commerce can offer.

Table 1 provides an overview of the current state of different generations of cellular voice and data services.

Increasing Bandwidth and Connectivity

Bandwidth has been doubling every 9 months, or roughly at twice the growth rate of computing power. Increasing bandwidth will lead to the creation of what

Table 1. Current state of different generations of cellular voice and data services (IC2 Institute, University of Texas, 2004)

Generation	Transmission Technology	Current Location
1G	AMPS (Advanced Mobile Phone Service)	U.S.A., but declining usage in metro areas
2G	CDMA (Code Division Multiple Access) TDMA (Time Division Multiple Access) GSM (Global System for Mobile Communications)	Mostly metro areas Being phased out Most of the world except U.S.A.
2.5G	GPRS (General Packet Radio Service) and CDMA 2000 1x	Current changes in the U.S.A. and some other areas
2.75G	EDGE (Enhanced Data Rates for Global Evolution)	In deployment phase in the U.S.A.
3G	CDMA2000 (Broadband CDMA) W-CDMA (Wideband CDMA)	Current push for use in the U.S.A. A standard in Japan and Europe

is called the "evernet," where billions of devices will be connected to the hyper-speed, broadband, multiformat Web. The high-speed networks of the 3G generation will provide additional capacity and enhanced functionalities. There is a strong need to combine the wireless (LAN; local-area network) concept and cell or base-station wide-area network design, and 4G is seen as the solution that will bridge the gap and therefore provide a much more robust network (IC2 Institute, University of Texas, 2004).

ISSUES AND CHALLENGES OF U-COMMERCE

U-commerce applications offer many benefits, but they also face challenges and raise new questions (Galanxhi-Janaqi & Nah, 2004). Mobile commerce faces the same problems troubling e-commerce plus a few of its own (Siau & Shen, 2003a, 2003b; Siau, Sheng, & Nah, 2003), and this is true for u-commerce, too. The higher value of u-commerce comes from the synergy created by its components. It is ironic that the same information practices that provide value to organizations also raise privacy concerns for individuals (Bloom, George, & Robert, 1994). The synergy between the u-commerce components increases the potential benefits, but it also adds new challenges. U-commerce inherits the privacy, trust, and security concerns of e-commerce, m-commerce, and other forms of digital commerce. Security and privacy are the two biggest concerns of consumers in embracing mobile commerce (Siau, Sheng, Nah, & Davis, 2004), and silent commerce applications, for example, may increase these concerns.

U-commerce emerges as a continuous, seamless stream of communication, content, and services exchanged among businesses, suppliers, employees, customers, and products (Accenture, 2001), and coordination becomes a fundamental issue. In addition, it is important for organizations to be able to determine the strategic directions of the organization, management's attitudes toward e-commerce initiatives, the potential of a learning environment (Kao & Decou, 2003), and how the evolution from e-commerce to u-commerce will be achieved. The u-commerce initiative will encompass a greater part of organizations' operations. As such, there will be a greater need for the coordination, synergy, and strategic integration of different initiatives (e.g., e-commerce, m-commerce, silent commerce, and others). Internet firms face serious political and legal uncertainties, which differ considerably between different markets (Frynas, 2002). Furthermore, in order to ensure long-term relationships and loyalty of customers, building trust is vital in u-commerce applications where you may never physically meet your customer. Building trust is a complex process that involves technology and business practices, as well as movement from initial trust formation to continuous trust development (Siau & Shen, 2003a). The companies that will be able to build trust in a u-commerce environment will be in a better competitive position.

The usability of devices is another concern, and future research is needed to improve the usability and various aspects of user interfaces (Nah & Davis, 2002). In addition, most of the devices in u-commerce are free from human intervention; this increases convenience on one hand but increases risks on the other hand (Galanxhi-Janaqi & Nah, 2004). New social issues arise as these u-commerce applications must mesh well with natural social behaviors or they will fail or lead to unforeseen outcomes (Grudin, 2002).

CONCLUSION

U-commerce will enable improved operating efficiency, enhanced customer services, increased service personalization, continuous supply-chain connectivity, and continuous interactivity. But there are still a number of obstacles that need to be overcome in order to fully realize the u-commerce vision. These obstacles include the lack of standardization, the difficulty of reading from and writing on very small devices, and privacy and security issues. In addition, culture and lifestyle are important factors that determine the adoption rate for u-commerce in different regions of the world. U-commerce will have broad implications for organizations themselves, as well. Companies involved in e-commerce still face obstacles such as the choice of business model, security and trust issues, integration with legacy systems, interoperability of systems with other organizations' systems, the assessment of the effectiveness of investments in technology, and the management of information overload. Furthermore, a comprehensive and unambiguous legal framework regarding online transactions is lacking.

Finally, the realization of this new vision will not be a replacement for other types of commerce, but an extension of them. U-commerce is the natural evolution of its components.

REFERENCES

Accenture. (2001). *The unexpected eEurope*. Retrieved December 2002, from *http://www.accenture.com/xdoc/en/ideas/eeurope2001/Full_Survey.pdf*

Bloom, P. N., George, R. M., & Robert, A. (1994). Avoiding misuse of information technologies: Legal and societal considerations. *Journal of Marketing, 58*(1), 98-110.

Frynas, J. G. (2002). The limits of globalization: Legal and political issues in e-commerce. *Management Decision, 40*(9), 871-880.

Galanxhi-Janaqi, H., & Nah, F. (2004). U-commerce: Emerging trends and research issues. *Industrial Management and Data Systems, 104*(9), 744-755.

Grudin, J. (2002). Group dynamics and ubiquitous computing. *Communications of ACM, 45*(12), 74-78.

IC2 Institute, University of Texas. (2004). *Austin's wireless future*. Retrieved June 2004, from *http://www.wirelessfuture.org/AustinsWirelessFuture.pdf*

Junglas, I. A., & Watson, R. T. (2003a). U-commerce: A conceptual extension of e-commerce and m-commerce. *Proceedings of the International Conference on Information Systems*, 667-677.

Junglas, I. A., & Watson, R. T. (2003b). U-commerce: An experimental investigation of ubiquity and uniqueness. *Proceedings of the International Conference on Information Systems,* 414-426.

Kao, D., & Decou, J. (2003). A strategy-based model for e-commerce planning. *Industrial Management and Data Systems, 103*(4), 238-253.

Nah, F., & Davis, S. (2002). HCI research issues in e-commerce. *Journal of Electronic Commerce Research, 3*(3), 98-113. Retrieved January 2004, from *http://www.csulb.edu/web/journals/jecr/issues/20023/paper1.pdf*

Nah, F., Siau, K., & Sheng, H. (2005). The value of mobile applications: A study on a public utility company. *Communications of the ACM, 48*(2), 85-90.

Schapp, S., & Cornelius R. D. (2001). *U-commerce: Leading the world of payments* (White paper). Visa International. Retrieved December 2002, from *http://www.corporate.visa.com/av/ucomm/u_whitepaper.pdf*

Siau, K., & Shen, Z. (2003a). Building customer trust in mobile commerce. *Communications of the ACM, 46*(4), 91-94.

Siau, K., & Shen, Z. (2003b). Mobile communications and mobile services. *International Journal of Mobile Communications, 1*(1/2), 3-14.

Siau, K., Sheng, H., & Nah, F. (2003). Development of a framework for trust in mobile commerce. *Proceedings of the Second Annual Workshop on HCI Research in MIS (HCI/MIS'03)*, 85-89. Extended abstract retrieved January 2004, from *http://cte.rockhurst.edu/sighci/icis_2003/HCI03_14.pdf*

Siau, K., Sheng, H., Nah, F., & Davis, S. (2004). A qualitative investigation on consumer trust in mobile commerce. *International Journal of Electronic Business, 2*(3), 283-300.

Watson, R. T., Pitt, L. F., Berthon, P., & Zinkhan, G. M. (2002). U-commerce: Expanding the universe of marketing. *Journal of the Academy of Marketing Science, 30*(4), 333-348.

KEY TERMS

Bluetooth: This is a low-power wireless-network standard that allows computers, peripherals, and consumer electronic devices to talk to each other at distances of up to 30 ft.

Electronic Commerce (E-Commerce): The conduct of business communication and transactions over networks and through computers. Electronic commerce is the buying and selling of goods and services, and the transfer of funds through digital communications.

Microelectromechanical Systems (MEMSs): These are chips that combine the capabilities of an RFID tag with small, embedded mechanical devices such as sensors.

Radio-Frequency Identification (RFID): The electromagnetic or electrostatic coupling in the RF portion of the electromagnetic spectrum is used to transmit signals. An RFID system consists of an antenna and a transceiver, which read the radio frequency and transfer the information to a processing device, and a transponder or tag, which is an integrated circuit containing the RF circuitry and information to be transmitted.

Silent Commerce (S-Commerce): The conduct of machine-to-machine transactions in real time without human involvement.

Television Commerce (T-Commerce): This is e-commerce occurring over the medium of television.

Ubiquitous Commerce (U-Commerce): According to Watson et al. (2002), it is "the use of ubiquitous networks to support personalized and uninterrupted communications and transactions between a firm and its various stakeholders to provide a level of value over, above, and beyond traditional commerce." It is the combination of electronic, wireless-mobile, television, voice, and silent commerce.

Voice Commerce (V-Commerce): The initiating of business transactions through voice commands.

Wireless Commerce (Mobile or M-Commerce): The buying and selling of goods and services through wireless handheld devices such as cellular phones and personal digital assistants (PDAs).

Understanding the Out-of-the-Box Experience

A. Lee Gilbert
Nanyang Business School, Singapore

INTRODUCTION

Appliances, ideally, are simple to use. You select one you think will meet your needs, get it home, take it out of the box, plug it in, and begin use. For a wireless home network, it may be necessary to read a manual, run a setup wizard, and make a few adjustments to get the performance you want, after which the new device delivers the performance its makers promised, every day, without fail. This scenario describes the expectations most people have from their first experience with networked mobile digital devices (NMD) such as laptop computers, home networks, personal digital assistants, and data-enabled cellphones. In many cases, their actual experiences are often far less satisfactory.

The personal computer, with the systems and applications software that enables our PC to perform useful tasks and entertain us, originated from computer science, a technical domain where few end users are in their comfort zone. Early PC users were confronted by vast manuals, then by cryptic error messages when things went wrong, which was often. Compare this user experience to acquiring and using a cordless telephone, which rarely requires reading a manual. Intended for use by everyone, the wireless network technology designed to link digitally enabled devices in our homes and offices is more complex than our phones. The design challenge is to ensure that, despite wide differences in their prior experience and intended use, each user has a satisfying experience with a new product.

There are good business reasons to confront this challenge. Early research in the diffusion of innovations (Rogers, 1965) posited that the first users of new technology seek to meet different needs and have different expectations for product performance than those that might later adopt the same innovation. Subsequent research across several product categories confirms this proposition and reveals that the sales growth for newly introduced products tends to "stall" if the skills set required for use are confined largely to members of the early adopter segment (Moore, 2003). As most costs to deploy new network-based services are essentially fixed, the rate of adoption often determines the difference between success and failure (Lucas 2003).

The computer industry was quick to realize the strategic role of the user experience. SRI and Xerox Parc pioneered user interface designs for personal computers, leading to today's Windows and Mac operating systems. In an era when Microsoft and other leading software firms adopted design standards such as "plug-and play" and in-context help, game developers such as Electronic Arts and Nintendo were designing and implementing applications that eliminated the help dialogue. Today, IBM promotes ease of use through its User Engineering and User Centered Design programs, and defines the out-of-the-box experience (OoBE) as "the initial experience a user has in taking a new product out of the box and setting it up, in preparation for use" (IBM 2004).

The home and office are not the only domains where OoBE management matters. Twenty-five percent of automobile components now involve some sort of computer; interactions between a car's information systems and its driver can be problematic. Stanford University's Center for Design Research is investigating how to design and manufacture cars to improve the driver experience (CDR 2004).

Despite these encouraging signs of progress, the consumer electronics industry is far from an ideal "5-Minute Ready" OoBE target (Bluez 2003). To get their PDA to work with an early Bluetooth dongle, users first determined installed software by entering the command: Use "ipkg status | grep bluez" and "ipkg status | grep rfcomm", then worked through about 20 instructions, and in some cases, entered 50 lines of new code. As a trip to CeBit will reveal, the industry plans to release a torrent of new digital products to link all digital devices at home and

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

in the office. While pervasive computing is appealing, it is unachievable, from a human factors perspective, without a new design philosophy.

BACKGROUND PERSPECTIVES ON THE OOBE

"Out-of-the-Box," a term of art with origins in industry, captures the shift in perspective from technical design, "what our product does," to refocus design effort on "how users experience our product during the critical first phase of use." Because digital products consist not only of a device, but software, instructions, and other arrangements for use (Rogers 1984), the "Box" metaphor is useful. For example, a GSM phone requires a Subscriber Identity Module issued by the network operator, while use of its data access capabilities may require settings both in the phone and at the network. The manual that accompanies the new phone may not contain all the information necessary to perform these tasks. In this case the user experience suffers because the "Box" is incomplete.

Modeling the OoBE

For any reasonably complex digital device such as a personal computer or smartphone, the OoBE results from a series of experiences over time rather than a single user event. These experiences are shaped by interactions among the form of the device, the user, and the user context. The form includes the physical design of the device and its accessories, its imbedded functionality, the accompanying "package" of complementary components such as software, arrangements for network access, and user documentation and manuals (Alexander 1977). The user context has both internal (motivation for use, abilities, etc.) and external (location, availability of suitable content, access to assistance, etc.) elements.

The use of each new device proceeds through five phases, beginning with selection, followed by acquisition, and first use, when users perform basic functions. Users expand the scope of use as they learn the capabilities of their new device, and may discover that it can meet emergent needs of which they were unaware. For example, new cellphone users may acquire their phones to make voice calls, learn to use text messaging from friends or family members, then move on to access an interactive service on the mobile Internet. When the use pattern matures, it stops expanding. The device may later be retired, either because user needs have outgrown device capabilities, or because the user no longer needs the services provided by the device. The OoBE mainly occurs across the first three phases of use, as portrayed by the "OoBE zone" arc in the model below.

Phase 1: Selection and Acquisition

Each OoBE is a function of the interactions among multiple entities over time. In most cases, a new user makes two decisions about the technology package: to acquire a given digital device, and to acquire or subscribe to a specific set of services. New NMD users base these interrelated decisions on the perceived fit among their expectations regarding the functionality of the device, the performance of the intelligent network to which it will connect, and the context in which they expect to use it. This complex set of perceptions and expectations shapes the context in which the user opens the box. This event may occur at the purchase point, in the office, at school, or at home.

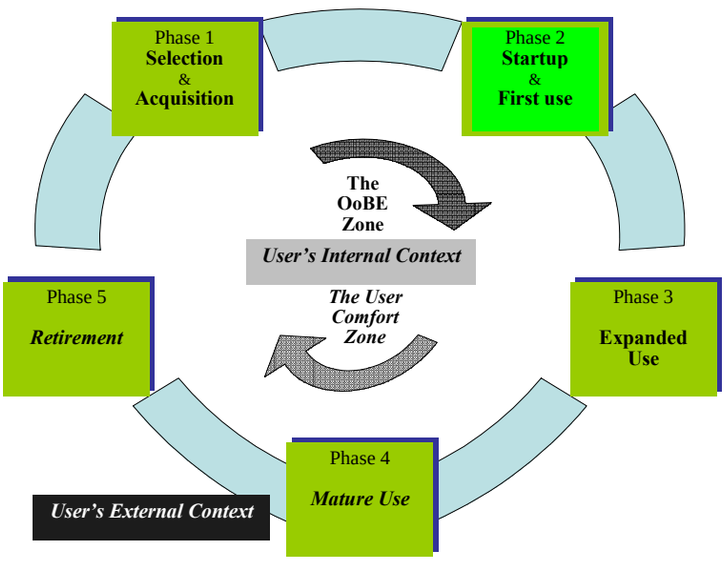
Phase 2: Startup and First Use

On opening the box, the user may find a system in one of several states:

1. The device, components needed for use, and detailed setup and use instructions.
2. The device, configured for immediate use, with simple "get started" instructions.
3. The device, missing one or more components needed for the intended function.
4. The device and related components, apparently not in working order.

The seller or provider may have inspected the package, configured the device, and pre-tested it prior to delivery. If the user opens the box at the purchase point, the seller may perform these services and provide basic instructions to the buyer on the spot. For a GSM phone with GPRS capabilities, a seller may have charged the battery prior to placing

Figure 1. A five-phase model of the OoBE life cycle



the box in stock, and can configure the phone and the user's SIM card and network account to provide access to the desired services at the time of delivery. This phase ends when a user has started up the new device, initiated the desired services, and successfully performed the primary functions for which the user acquired the system. There are, however, many opportunities for failure at this phase. During the initial rollout of WAP services, few sellers were they able to assist new users to configure the system (phone, SIM card, and network) properly, nor did they understood how to access content once the system was ready. As there were few WAP users, it was difficult to find a colleague or friend to help. As a result, many users abandoned their desire for WAP services without ever having accessed them. When this phase goes wrong, some users put the new device back in the box and revert to their old one if they still have it, while others buy a replacement.

Phase 3: Expanded use

The flexibility inherent in programmable digital devices means that we can acquire them not only to satisfy our primary needs, but to "see what they can do" with respect to other needs. Once they succeed in meeting their initial needs, some users go back to the box, get out the directions, and experiment with the system to discover how it might meet other needs. Some new communications capabilities have the potential to generate rapid growth in use, as seen in the runaway successes of mobile text and instant messaging. Others, such as WAP, were initially less successful, but gradually attracted an increasing number of users. As this user community grows, it has the capacity to assist others, either in person or through online forums and chat rooms dedicated to this activity. Realizing that there are many opportunities for failure, some operators have developed Web sites to support expanded use of these devices. Access to such resources can be "included in the box", in the form of the information or training material required to access and use them.

ANALYSIS OF THE OOBE LIFE CYCLE

The complete OoBE takes place across three phases in the proposed model. The first, selection and acquisition, is crucial because this process establishes user expectations of the experience to follow. Note that the IBM Ease-of-Use perspective on the OoBE focuses entirely on initial setup and use, or phase 2 (IBM, 2004). During the third phase, ex-

panded use, the user's ability can develop sufficiently to move beyond initial (and often trivial) first use applications. Combined, the three processes frame user perceptions of the experience and define the content of favorable (or otherwise) word-of-mouth communications that may encourage or discourage others to adopt.

In the table above, activities by the entities that determine the OoBE are mapped to phases for the case of a data enabled cell phone. The complete OoBE (green shading in the table) is a dynamic phenomenon that begins with device/network service selection and ends as users enter their "comfort zone," when their use behavior reaches maturity. The model reveals that the success of OoBE design is determined partly by the device itself, and partly by interactions among other entities. The role of each entity shifts with each phase in the life cycle. The user and the user context(s) are indivisible elements of the OoBE experience. As users have different motivations and skills, and their use occurs in varying contexts (Prekop and Burnett 2003), identically configured terminal/network packages may generate highly contrasting OoBE for individual users. This implies that a "one size fits all" OoBE design solution is unlikely to succeed for all users.

As it is impossible to design a package for each individual user, it is necessary to sort users into groups that share attributes relevant to the OoBE design task. Research by the Information Management Research Centre (IMARC) at Nanyang Business School in Singapore suggests that market segmentation techniques (Weinstein 1994) are appropriate for this task. Studies of mobile data services (MDS) usage patterns for several hundred NMD users led to a dynamic segmentation model based on the time of adoption, information channels that influenced user decisions, and primary needs met through use (Gilbert and Han 2004). The segments, ordered by mean time of MDS adoption, are:

- **TechnoToy:** Filling needs for hands-on knowledge about technological developments.
- **Mobile Professionals:** These services create new value related to work life, including calendaring, and access to mobile e-mail and intranet/extranet services.
- **Sophisticates:** Filling needs for status, in terms of material style.
- **Socialites:** Filling needs to keep in touch with family and friends while on the go.
- **Lifestyle:** These services, partly overlapping the categories listed above, fill convenience needs related to mobile lifestyles, such as delivering information or directions to people who are in an unfamiliar location, and helping people fill "dead time" with time-critical tasks. Examples of such tasks include bill paying while waiting in line or on public transport, or facilitating meetings among friends who are on the move.

These five segments contained those most likely to adopt MDS. Two added segments, containing users whose needs were unlikely to motivate them to adopt, emerged:

Table 1. OoBE across the life cycle for a Networked Mobile Digital Device (NMD)

	Life Cycle Phase for NMD Use				
Entity	Selection	Startup/ 1st use	Expanded use	Mature use	Retirement
Mobile terminal	Packaging Reputation	Basic UI design Setup process	Ease of Use Consistent UI	Performance matches needs	Performance lags user needs
Intelligent network services	Promised MDS access and performance	Configuration Performance User support	Access to new MDS, roaming User support	Value-added services, QOS, User support	MDS use exceeds system capabilities
Internal MDS user context	Motivation to acquire MDS device, skills	**Core OoBE experience**	Emergent need to expand service scope	Needs match service / device capabilities	Desire more service / device performance
Social context	Opinion leadership	Expanding user community	Support for new uses	Sustained support for use	Communicate emerging uses

- **Misers:** Members of this segment were unwilling to pay for wireless data services.
- **Laggards:** Were the last to know about and adopt new technologies.

Focus group discussions revealed that only a minority of users were able to configure their NMD to access MDS without assistance, but that many abandoned MDS after an initial period of experimentation. This finding contrasts sharply with the runaway success of i-Mode in Japan, resulting from an intensive design focus on the user experience (Matsunaga 2000). However, MDS users from early adopter segments were somewhat less likely to seek assistance, and far more likely to continue MDS use.[1]

FUTURE TRENDS IN MANAGING THE OOBE

In the sense that many types of digital devices with communications capability are available, at least in a primitive form, the future is here. Soon to follow is exponential growth in the adoption of such devices, and an attendant increase in the number and complexity of wireless networks that link them. To illustrate this complexity, count the number of digital devices that are already present in homes and offices, then estimate the combinations and permutations for network design. Then, identify the many vendors that make devices and the contrasting contexts from which they originate, and consider the potential conflicts in network standards, user interface design, and device packaging.

The table above draws on IMARC research that sought to understand why cellphone subscribers in Singapore, where many environmental conditions favored rapid diffusion, were slow to adopt digital MDS (Chia, Hazam, & Ho, 2001). The OoBE model captures many factors that act as barriers to diffusion and use, and informs product designers, marketers, and other stakeholders that shape the OoBE (Urban and Hauser, 1980).

AN AGENDA FOR OOBE RESEARCH

Human factors will come into play (Chapanis 1996). Context-aware agents may help reduce the information overload caused by mismatches between increasing content and the small displays and weak processors in mobile devices (Lee & Lu 2003). Agents may also help users rapidly adapt to another context, in response to environmental turbulence (Woods, 1988). Such context aware applications will need to focus not only on external contextual factors such as location content availability, and access to the user community, but to incorporate the cognitive domain by including internal factors (Prekop & Burnett, 2003). Internal contextual factors may include the goals, tasks, work processes, personal events, emotional and physical state, and knowledge or skills of users (Rosen, Purinton, & Lloyd, 2003).

Table 2. Barriers to the diffusion of MDS use, user context, and OoBE phase

Barriers to diffusion of mobile digital device	Context	Design factors	Phase
1. Life cycle cost to acquire and use device	External	Value Design, pricing	1
2. Lack of compelling service content	Internal	Value Proposition	1, 2, 3
3. Physical limitations of mobile devices	External	Device, Network, and Applications design, links to other devices	2, 3
4. Risks inherent in use (security, etc)	Internal	Applications and Infrastructure design	2, 3
5. Inadequate network performance -	External	Network and database design	2, 3
6. Lack of user knowledge and/or skills	Internal	Packaging, Device, User Interface, and Applications design	1, 2, 3
7. Conflict with social or organizational norms	External	Packaging, Device, User Interface, and Applications design	2, 3

Table 3. Current knowledge gaps in OoBE Design, by phase

(phase) **Knowledge Gaps**	**Promising Areas for Research**
0. Currently, each provider configures one or more "packages," composed of the digital device, accessories, and arrangements for use, to meet a range of user needs.	o Patterns to describe "fit" between user contexts and package configurations, o Model to segment user community into a relatively manageable number of groups
1. Users select the package they believe will meet those needs of which they are aware.	o Agents to help users define their needs, select, configure, and use a new package.
2. The user interacts with this package in a specific context or contexts, over time. 2. Either the external or internal context may change, sometimes very rapidly, during use 2. During initial use of the package, the user may (or may not) satisfy the initial needs set.	o Design principles for context-aware packages that evolve with users. o Context-aware packages that respond to changes in the user environment o Design principles for user tracking and diagnosis to monitor OoBE quality and initiate action to remedy shortfalls.
3. During in-context interaction with the package, the user may discover emergent needs, and learn how to meet them.	o Design principles for the capture and dissemination of information about new uses to the user community

As the crossroads where behavior meets technology is inherently interdisciplinary, the search for design solutions extends beyond the design and human factors communities. While anthropologists and other social scientists study human activity and can contribute to certain aspects of design, planners and architects also offer deep insights into the interactions among people and man-made artifacts (Kaplan, 1973). A design language developed for architects (Alexander, 1977) was adapted to software (Scanlon, 2004), and the design of mobile applications (Roth, 2002). The advantage of a pattern language for design is that it captures the logic of a given context and facilitates re-use of building blocks (ranging from general design principles to actual computer code) for solutions.

An understanding of the contexts that motivate use is essential: while most NDD uses (e.g, SMS and IM) are deeply collective in nature, others (investment transactions and games played privately on the NDD) deliver benefits to individual users. Table 3 maps promising areas for OoBE research to the first three phases of the proposed model.

CONCLUSION

The promise of ubiquitous computing is fulfillable only if the design community creates a satisfactory experience for each set of adopters, who will in turn encourage those that follow. This will demand a new approach to the design and implementation of the "out-of-the-box" experience. This article establishes a requirement to expand our perspective of the out-of-the-box experience beyond initial use, to encompass the selection and acquisition process during which users form expectations, plus the period after first use, during which the scope of use expands to fulfill emergent needs. It proposes an expanded life cycle model to capture activities by all the entities involved throughout the life cycle, not only the user. It also makes a case for research to define segmentation, based on the user context and skill set, and the development of OoBE designs that are suitable for each user segment. Analysis of this and other research opportunities highlighted by the model strongly suggests the need for an interdisciplinary and user-centered approach to design.

REFERENCES

Alexander, C. (1977). *A pattern language.* London: Oxford University Press.

Bluez (2003), text message in online Bluetooth discussion group, accessed March 15, 2004, at *http://bluez.sourceforge.net/download/zaurus/README.bluez.zaurus.new.txt*

Chapanis, A. (1996). *Human factors in systems engineering*. New York: John Wiley & Sons.

Chia, C.H., Hazam A.,,& Ho, K.A. (2001). WAP: The impact of the mobility mandate in Asia, a

Singapore perspective. Research Report, Singapore: Information Management Research Centre.

Davis, D.M. (1993). Social impact of cellular telephone usage in Hawaii. *Proceedings of the Pacific Telecommunications Conference*. Honolulu: Pacific Telecommunications Council, 641-648.

Gilbert, A.L. & Han, H. (2004). Modeling the dynamics of emerging mobile data services markets, in N.S. Shi (Ed.) *Mobile commerce applications.* Hershey, PA: Idea Group, 233-255.

IBM (2004). Designing the out-of-box experience, accessed March 15, 2004, at *http://www-306.ibm.com/ibm/easy/eou_ext.nsf/Publish/626*

Jokela, T. (2002). A method-independent process model of user-centered design. *Proceedings of IFIP World Computer Conference 2002*: Montreal, 23-28.

Kaplan, R. (1973). Predictors of environmental preference: Designers and clients. In W. Prieser (Ed.) *Environmental design research*, Vol.1, Stroudsberg: Dowden, Hutchinson and Ross.

Lee, W.P. & Lu C.C. (2003). Customising WAP-based information services on mobile networks, *Personal and Ubiquitous Computing,* (7), 321-330.

Lucas, H. (2003). *Strategies for electronic commerce and the Internet.* Cambridge: MIT Press.

Matsunaga, M. (2002). *The birth of i-Mode*. Singapore: ChuangYi Press.

Moore, G. & McKenna, R. (1999). *Crossing the chasm: Marketing and selling high-tech products to mainstream customers.* New York: HarperBusiness.

Prekop, P. & Burnett, M. (2003). Activities, context, and ubiquitous computing. *Computer Communications* (26), 1168-1176.

Rogers, E. & Kim, J.I. (1984). *Communication technology.* New York: The Free Press.

Rosen, D., Purinton, E., & Lloyd, S. (2004). Web site design: Building a cognitive framework. *Journal of Electronic Commerce in Organizations, 2*(1), 15-28.

Roth, J. (2002). Patterns of mobile interaction. *Personal and Ubiquitous Computing,* 6, 282-289.

Urban, G. & Hauser, J. (1980). *Design and marketing of new products.* New York: Prentice Hall.

Weinstein, A. (1994). *Market segmentation*. Chicago: Probus.

Woods, D. (1988). Coping with complexity: The psychology of human behaviour in complex systems. In L. Goodstein, H. Andersen, & S. Olsen (Eds.), *Tasks, errors, and mental models.* London: Taylor and Francis, 128-148.

KEY TERMS

Global Systems Mobile (GSM): Industry standard for second-generation digital cellular communications networks, soon to be superseded by Third Generation (3G) networks.

IMARC: Information Management Research Centre at Nanyang Business School, Singapore.

i-Mode: Brand name for voice plus a wide range of data services delivered by NTT Docomo in Japan.

Instant Messaging (IM): Applications that provide immediate delivery of messages over fixed-line and mobile IP networks.

Mobile Data Services (MDS): Delivery of content, for example, music to MP3 players or games to consoles, over wireless networks.

Networked Mobile Digital Devices (NMD): Network-enabled laptop computers, personal digital assistants, wireless home networks, cellphones, and wireless portable music players and game consoles.

Out-of-the-Box Experience (OoBE): User perceptions, relative to prior expectations, of their early use experience with a new object, including acquisition, setup, initial, and subsequent use of the object in a specific context.

Personal Digital Assistant (PDA): Portable computing device that provides computing and information storage and retrieval capabilities for personal or business use. Lacking a full keyboard, the PDA fits the palm.

Short Messaging Service (SMS): Text-messaging service over GSM networks.

Smartphone: Mobile phone with the capabilities of a PDA plus wireless access to cellular networks.

Subscriber Identity Module (SIM): Smart card deployed in the mobile terminal, used by GSM network to identify the subscriber.

ENDNOTE

[1] Although supporting data for this analysis draws on a single market (Singapore), and must be treated with caution, the island nation has the potential to serve as a lead market for an unwired community.

A Unified Information Security Management Plan

Mari W. Buche
Michigan Technological University, USA

Chelley Vician
Michigan Technological University, USA

INTRODUCTION

Information is quickly becoming the most significant asset of business practice, and it must be protected and secured in order to be useful. Information security, intrusion detection, and privacy were in the top-10-issues list from the American Institute of Certified Public Accountants (AICPA) survey ("Information Security Heads Top 10," 2003). Furthermore, the potential severity of attacks encourages collaboration between vendors and business clients, including educational institutions, in combating threats (Cox & Kistner, 2003). Essentially, transactions generate data that must be stored for future access in the form of information[1]. Therefore, data integrity is essential because it directly impacts information quality and any decisions based on that information (Brogan & Krupin, 2003; Ross, Stoneburner, Katzke, Johnson, & Swanson, 2003).

To increase confidence in information quality, the information and data must be secured from threats. Information security management must address each of the key areas—confidentiality, authentication, authorization, data integrity, and nonrepudiation—while allowing for continued optimal performance (Gurski, 2003). A rather extreme alternative would be to only distribute needed information to particular individuals, eliminating the necessity of open access to information storage devices (Brogan & Krupin, 2003). However, as the business community strives for integration and sharing of data resources, the criticality of information protection increases, making this an important topic for information systems personnel and systems users (Whitman & Mattord, 2003).

This paper addresses information security management concerns and is divided into five major sections. The first section presents a general discussion of the history of information security management, extending back to the roots of computer security. The next section identifies the key components of information security management, including software, hardware, and human and social elements. This is followed by future trends and concerns relevant to the topic of information security management. Finally, we present our conclusions and general implications for practitioners and academic researchers.

BACKGROUND

Today's discipline of information security has evolved from the early management efforts referred to as computer security (Whitman & Mattord, 2003). Computer security focused on safeguarding physical computing devices and output. Soon after the introduction of computers in office environments, it became apparent that a method for managing the hardware was needed. Locking the office door was usually sufficient for securing the physical computer equipment, as the machines were large and not easily transported. These early computers were initially intended to automate clerical processes, so the actual information generated was not viewed as highly sensitive. Output could be controlled and regulated like all other sensitive artifacts since the output was often in the form of paper printouts.

The lack of integration of the machines contributed to the simplicity of early computer security management (Whitman & Mattord, 2003). Before file sharing became commonplace, files resided in only one location and could be controlled using passwords with relative assurance of protection. Ownership and possession of information could be

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

easily identified and managed. The advent and ubiquitous nature of the Internet has intensified and complicated the management of information security since no organization controls or manages the vast network of networks (Dhillon, 2003; United States General Accounting Office, 2004; Vijayan, 2003; Weiss, 2004). However, the security of each connected device directly affects the security of every other machine and peripheral on the network. So, management is justified in asserting that security is everyone's business, extending the responsibility beyond security management personnel (Parker, 2003; Verton, 2004).

Legislation is also driving improvements in information security (Johnson, 2004). Two particularly complex regulations pertaining to information technology are the Health Insurance Portability and Accountability Act (HIPAA; Brewin, 2003) and the Sarbanes-Oxley Act in the United States. Both laws make organizations responsible for the protection and control of personal information of patients, customers, and stakeholders. The HIPAA standards have recently been revised to place the burden of risk assessment regarding patient information on the reporting agency. This action decreases the mandated provisions of compliance, but raises the level of accountability for IT managers (Brewin, 2003). The Sarbanes-Oxley Act also places the accountability and responsibility for compliance with the firms' executives. Essentially, business practices need to be reorganized to emphasize privacy and security, embedding security in the processes rather than reacting to breaches in an ad hoc manner (Dhillon, 2003; Gurski, 2003). The legislation is intended to strengthen consumer trust and forge stronger relationships between firms and stakeholders.

INFORMATION SECURITY MANAGEMENT PLAN

Information security involves establishing policies and procedures intended to prevent and/or detect unauthorized intrusions into the organization's information system (Ross et al., 2003; Whitman & Mattord, 2003). Whenever possible, management should combine multiple layers of security to ensure adequate coverage. At the same time, the plan needs to be unified (Johnson, 2004; Myers, 2003; Whitman & Mattord). That is, the separate elements of the information security plan must be managed as a single effort or strategy (Swartz, 2004; U. S. GAO, 2004; Weiss, 2004). As part of the firm's IS strategy, management should perform a risk analysis and determine the appropriate level of security required based on a variety of factors such as possible threats to the network, anticipated damages from expected threats, the probability or likelihood of the threat occurring, and the impact of the information security plan on the organization (e.g., the culture; Whitman & Mattord, 2003). A successful information security management plan must contain policies and procedures that cover multiple dimensions: hardware, software, and people (employees) are the three primary components, as depicted in Figure 1. The major aspects of an information security plan are summarized in Table 1, which provides nine types of barriers to unauthorized access. First, physical security involves the protection of tangible assets by preventing actual access. For example, the computer servers and electronic equipment are secured by preventing physical entry, locking doors, or affixing computers to immovable structures. This is one of the most basic solutions to securing hardware, but is often overlooked in the creation of an information security plan (Berti, 2003). Hackers claim that if they can physically touch a computer, they can hack into the system (Crume, 2000).

Second, personnel controls include the procedures used to limit the access of employees through the use of passwords and security profiles (Jamieson & Handzic, 2003). Passwords must be sufficiently rigorous so that hackers cannot easily break them (Cox & Kistner, 2003; Crume, 2000). To reduce the burden of remembering so many passwords, workers often resort to documenting the codes and placing them near their workstations. Should an unauthorized person gain physical access to the workstation, the list of passwords would allow the individual easy entry into the system. Likewise, security profiles should be established to restrict employee access to only necessary information required in the performance of the job based on the need to know. Exceptions can then be managed on a case-by-case basis, allowing greater access as required to fulfill specific duties.

Third, system software provides security embedded directly in the operating system (Whitman & Mattord, 2003). In most instances, however, the vendor default settings do not provide adequate security. Network administrators or security specialists must reconfigure the software during installation to set the security to the desired level of protection.

Fourth, application software is functionally designed to create barriers to entry and/or detect probable intrusion. There are many vendors creating security software. Companies should select software that provides the desired security at an acceptable price.

Fifth, service continuity and data recovery are sometimes not included in information security plans. It is important to ensure the recovery of business processes following an intrusion, such as a denial-of-service attack. In the midst of a security crisis, the plan should detail proper actions to mitigate the impact of the threat, preserve evidence for later evaluation, and restore the system to regular operation (Ross et al., 2003). Extended periods of IS failure can lead to millions in lost revenue for a company.

Sixth, monitoring and auditing include regular assessments of the performance of the plan. At least on a yearly basis and whenever significant changes are made to information systems, the plan must be tested (Ross et al., 2003; Whitman & Mattord, 2003). Without vigilant security personnel, holes in security will go undetected for extended periods of time. To be most effective, feedback from assessments is then used to make improvements to the existing plan.

Seventh, information security must be driven by top executives (Berti, 2003). Without this level of support, employees are less likely to incorporate or adopt security policies. They will view security as additional work that is not valued by the company, leading to complacency and disregard for procedures. Management needs to implement a plan to motivate employees to incorporate security into their daily work routines (Parker, 2003).

Eighth, direct supervision of information security by trained professionals should not be minimized (Verton, 2004). Johnson (2003) recommends that specialists in this field should pursue certifications based on knowledge attainment, earn a graduate degree in information security, develop disaster-recovery and risk-management skills, build a home laboratory to practice these skills, contribute to the community, work with strategic partners to acquire other perspectives, seek opportunities to work in information security, and consider public service. Strong and diverse experiences within the career field will enhance the skill set of the security professional, benefiting the organization.

Finally, training is one of the most essential elements of the information security management plan. Employees must understand the policies and procedures of information security. Workers must also be familiar with management expectations with regard to privacy rules. Databases and systems contain highly sensitive information, and employees must accept responsibility for the safekeeping of that information. Individual pieces of information might not seem significant; but, when related pieces are combined, business secrets could inadvertently be revealed. Therefore, increasing employee awareness of individual responsibilities in information security is the primary function of training.

Components of Information Security

In order to accomplish the objectives of an information security management plan, managers establish controls designed to block intrusion from unauthorized users within the company (internal environment) and hackers with malicious intent attempting to enter the systems via Internet connections (external environment). The major components of information security management are hardware, software, and employees (i.e., humans). As Figure 1 shows, the three components are integrated to defend against attacks from various threats. Each of the three individual components is discussed below.

Hardware

The physical equipment used in the development of information systems should contain at least a minimal amount of built-in security. Firewalls can be either part of the hardware or a software application. As previously discussed, physical security is necessary to prevent basic violations. Many benefits of computers tend to create additional security concerns. For instance, the desirability of laptops is their portability. However, the convenience of the

Table 1. Types of barriers to be included in an information security plan

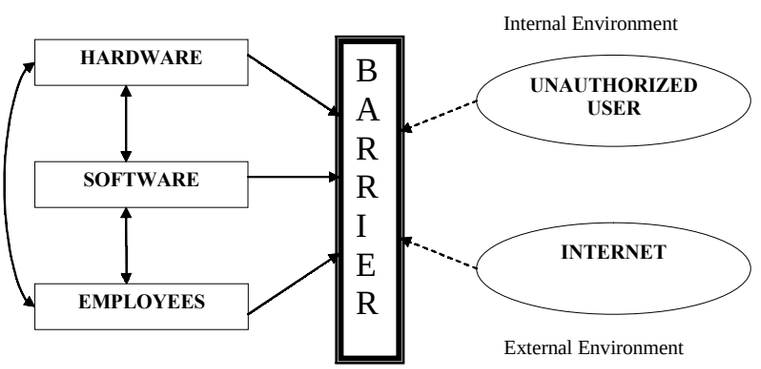
ELEMENT	COMPONENTS PRIMARILY AFFECTED	DESCRIPTION
Physical Security	Hardware, Software, & Employees	Preventing unauthorized access to tangible assets
Personnel Controls	Software & Employees	Actions taken to ensure employees are restricted to accessing only information required for the completion of their work
Systems Software	Software	Security practices are embedded in the operating system.
Application Software	Software	Software created to provide barriers to entry and intrusion detection
Service Continuity	Hardware, Software, & Employees	Ensuring the recovery of business processes following an intrusion such as a denial-of-service attack
Monitoring & Auditing	Employees	Regular assessments of the performance of the plan. Feedback is then used to improve on the existing plan.
Management Support	Employees	Information security must be driven by top executives.
Security Expertise	Employees	Trained professional in charge of information security
Training	Employees	Education of employees regarding corporate policies on information security

light weight and compact size of laptops causes the hardware to become more vulnerable. Alternative solutions to physical security must be developed to respond to evolutionary changes and trends in technology. Innovations create additional threats that must be dealt with in a proactive fashion.

Software

Firewalls and antivirus software assist computer users by automatically creating a barrier to prevent unauthorized entry. Various types of software can be used to block malicious attacks or for intrusion detection. Even if the information was not stolen or modified, it is important for companies to know when their systems have been breached. Intrusion detection systems (IDSs) perform this function of maintaining vigilance at all times through network monitoring (Schneier, 2000). They basically look for any type of suspicious behavior and report those observations (i.e., evidence of an attack or an attack in progress) to the system administrator through the

Figure 1. Components of the information security management plan



use of a security log. Good IDSs are accurate in their assessments, timely in reporting the attacks, and provide diagnostic suggestions on how to respond to the attack (Schneier, 2000).

Other methods used to protect data are encryption and digital signatures. Encryption of data is beneficial because even if unauthorized users gain access to data, the results are garbled and useless nonsense. Only intended recipients would have the ability to decrypt the data, restoring it to a useful format. Confidentiality can be assured because the message is not intelligible to unauthorized users. Confidentiality means that information is protected from interception by unintended parties, thereby preventing disclosure (Panko, 2004; Whitman & Mattord, 2003). Similarly, digital signatures permit senders and receivers of information to pass messages that can only be viewed by an entity possessing the appropriate key for decoding the message. These methods reduce the need for securing the channels that messages travel through because the message is indecipherable to all but the intended recipient. In addition, digital signatures and encryption techniques decrease the likelihood that one party in a transaction can deny involvement or participation. This is a legal decision called nonrepudiation (Crume, 2000). Nonrepudiation implies that the actions accomplished based on the acceptance of a private key are binding to the holder of that private key (Crume, 2000). Therefore, it is critical for individuals to protect their private keys.

Another concern that is becoming more common is the use of patches to keep software current. That is, patches are created by vendors to close security holes that could be exploited by hackers. Without a procedure for routinely scanning for and applying patches, corporate personnel might be relying on faulty security software, thereby creating a false sense of security (Crume, 2000).

Employees or Humans

Humans create many of the vulnerabilities that put companies at risk. As discussed earlier, training is necessary to raise awareness and to teach appropriate responses to threats. A specific threat that is very common is called social engineering (Berti, 2003; Mitnick, 2003). This approach, used by unauthorized hackers, exploits human behavior. In particular, hackers seek to gain entry to buildings and/or information systems through seemingly innocent requests for assistance. Since the basic human tendency is for people to be helpful, employees often unwittingly provide sensitive information without considering the potential consequences. For example, a person carrying a cumbersome box approaches a door that requires an identification-card swipe to gain entry. All that might be necessary is a pleading look at a legitimate employee, and the unknown person is granted access without ever showing proper identification. Persons with malicious intent will look for easy targets before launching their attack (Berti, 2003). This situation is compounded by the lack of personal contact with customers and other employees in large, global enterprises. Employees might never meet their colleagues, supervisors, or customers face to face. Therefore, the current dependence on electronic communications actually increases the vulnerability of employees to social engineering. Social engineering can be performed over the telephone to obtain passwords or access to information systems, physically as in the example of gaining access to a building or controlled area, or electronically through mass e-mail messaging. The greatest danger from social engineering is that it can effectively bypass any technological and procedural barriers an organization develops (Scambray, McClure, & Kurtz, 2001). For these reasons, employees are often considered to be the weakest link in the security plan (Schneier, 2000).

The best defense against social engineering is training. Trustworthy employees do not intentionally violate company security policies, but they often become complacent with regard to daily security procedures. Ideally, training should consist of exposure to a number of different scenarios so employees learn to recognize common tactics of social engineering. Biometric readers in combination with passwords would also help to reduce the damage realized by the sharing of information (Schneier, 2000). Essentially, the password communicated would be worthless without the additional biometric information.

Section Summary

This section provided details about the various security controls (or safeguards) that should be included

in an information security plan (Gurski, 2003; Ross et al., 2003). Although it is often easier to implement changes in modules, the only way to optimize security is by creating an integrated plan. The individual components of hardware, software, and employees must work in unity, not as individual silos of protection (Ross et al.; U. S. GAO, 2004). Also, it is imperative that everyone in the organization contributes to the goal of security by being vigilant, from the top executives to frontline staff. Information security must be a philosophical change, penetrating every aspect of the organization and becoming part of the corporate culture.

FUTURE TRENDS

Internet usage and the integration of computer networks continue to expand over time. Sharing files of all types of information facilitates group decision making through improved communication, coordination, and cooperation. Geographic distance is no longer a barrier for business or social units. The very benefits of computing networks create system vulnerabilities. How can a company protect its systems without giving up the flexibility of networks? That is the goal of information security management. Yet, according to the media, new incidents of security violations commonly occur. Selected examples include identity theft, denial-of-service attacks, and the remote control of personal computers (PCs). These violations affect consumers and commercial ventures, public and private entities, and shake individual confidence in computing solutions.

Identity theft is one of the primary concerns of law-enforcement agencies and companies with online access (Morrell & Kroen, 2003). Consumers, particularly those using the Internet for purchasing goods and services, are cautioned to use care when providing sensitive personal data (i.e., social security numbers, credit card information). Electronic commerce has experienced slower growth than originally expected due to apprehension and consumer anxiety about completing transactions over insecure Web sites. Each publicized violation of Internet security produces a ripple effect; consumers become more wary and vendors become more proactive in promoting the security of their own sites. Information security is as important as price comparisons to many savvy online shoppers. Privacy concerns should not be considered overhead expenses; attention to these issues can actually contribute to "brand image and trust" (Gurski, 2003, p. 5).

Denial-of-service attacks also impact e-commerce sites. The hacker often takes advantage of known weaknesses in the software code, creating a buffer overflow. A buffer overflow takes advantage of poorly written code by inputting additional characters that allow a hacker to input false code for execution (Schneier, 2000). The characters confuse the computer into reading the extra portion of the message as a new command, potentially creating the opportunity for the hacker to gain control over the system. Protection against buffer-overflow vulnerabilities is accomplished through careful programming, imposing restrictions on the length of characters, and not allowing additional characters to overwrite existing code. In some cases, denial-of-service attacks have caused servers to lock up by sending an excessive number of messages to the site. Consumers will give up on a site, choosing alternatives to complete their shopping transactions. Many large firms have experienced costly attacks of this nature (Whitman & Mattord, 2003).

Broadband Internet connections have improved the speed of access for users. The negative aspect is when the PC user maintains an open connection to the Internet while not actively using the machine (Piazza, 2003). The open gateway creates a vulnerability that is very tempting to hackers. Software (e.g., PC Anywhere) that enables employees to remotely access their personal computers also provides that same capability to hackers. While the application can facilitate telecommuters working at home and accessing files at work, the increased risk must be addressed by information security management.

Wireless local-area networks (WLANs) create additional security requirements (Myers, 2003). New protocols are being implemented that provide greater protection of transmissions than previous formats, using longer keys and more advanced algorithms (Myers, 2003). The challenge for security professionals is to stay ahead of the hackers since both groups continue to evolve and innovate.

An innovative solution that is gaining support is biometrics (Morrell & Kroen, 2003). Biometrics

involves using a physical identifier that is unique for each person. Examples in practice are fingerprints and retinal or iris scans. The drawback is expense; the equipment is currently cost prohibitive for all but large corporations, but prices are decreasing.

As a natural consequence to the rapidly increasing collection of data on security violations, vendors are creating "security event management" software to analyze the data (Vijayan, 2004). Essentially, managers need to collect and assess their companies' information vulnerabilities efficiently and effectively in order to respond in a timely manner.

CONCLUSION

Information security is not an easy process to manage. However, the dire consequences to businesses when they neglect security cause significant attention to be directed at the problem. Management cannot afford to take a casual approach to security. In fact, continuous monitoring and updating of both hardware and software are the keys to successful security. The other major component of the security triad is employee training. Instruction must be required, comprehensive, and recurring. Complacency can lead to catastrophic outcomes, providing opportunities for unauthorized entry. With the present importance placed on information and data integrity, information security must remain at the very top of management's list of priorities.

REFERENCES

Berti, J. (2003). Social engineering: The forgotten risk. *Canadian Hr Reporter, 16*(13), 21-23.

Brewin, B. (2003). HIPAA data rules leave choices to IT. *Computerworld, 37*(8), 1, 16.

Brogan, J., & Krupin, P. J. (2003). Access denied. *Industrial Engineer, 35*(10), 40-43.

Cox, J., & Kistner, T. (2003). Security lesson. *Network World, 20*(27), i-ii.

Crume, J. (2000). *Inside Internet security: What hackers don't want you to know.* London: Addison-Wesley.

Dhillon, G. (2003). Data and information security. *Journal of Database Management, 14*(2), 1.

Gurski, M. (2003). *The security-privacy paradox: Issues, misconceptions, and strategies.* Toronto, Ontario, Canada: Information and Privacy Commissioner and Deloitte & Touche.

Information security heads top 10. (2003). *The Practical Accountant, 36*(3), 18.

Jamieson, R., & Handzic, M. (2003). Impact of managerial controls on the conduct of KM in organisations. In C. W. Holsapple (Ed.), *Handbook on knowledge management* (chap. 25, pp. 477-505). Berlin, Germany: Springer.

Johnson, A. H. (2003). Boost your security career. *Computerworld, 37*(28), 41.

Johnson, M. (2004). Compliance bonanza. *Computerworld.* Retrieved May 28, 2004, from *http://computerworld.com/governmenttopics/government/policy/story/0,10801,92961,00.html*

Mitnick, K.D. (2003). Best practices: Are you the weak link? *Harvard Business Review, 81*(4), 18.

Morrell, J., & Kroen, T. (2003). Top tech trends. *Credit Union Magazine, 69*(12), 4A-10A.

Myers, R. (2003). Combine VPN and encryption. *Communications News, 40*(11), 34.

Panko, R. R. (2004). *Corporate computer and network security.* Upper Saddle River, NJ: Pearson/Prentice Hall.

Parker, D. B. (2003). Motivating the workforce to support security. *Risk Management, 50*(7), 16-19.

Piazza, P. (2003). Information security perception gap. *Security Management, 47*(9), 50.

Ross, R., Stoneburner, G., Katzke, S., Johnson, A., & Swanson, M. (2003). *Recommended security controls for federal information systems* (NIST special publication 800-53). National Institute of Standards and Technology, Computer Security Division, Gaithersburg, MD.

Scambray, J., McClure, S., & Kurtz, G. (2001). *Hacking exposed: Network security secrets & solutions* (2nd ed.). Berkeley, CA: Osborne/McGraw-Hill.

Schneier, B. (2000). *Secrets & lies: Digital security in a networked world.* New York: John Wiley & Sons.

Swartz, N. (2004). Survey assesses the state of information security worldwide. *Information Management Journal, 38*(1), 16.

United States General Accounting Office (GAO). (2004). *Information security: Further efforts needed to address serious weaknesses at USDA* (GAO-04-154 bulletin). Retrieved May 28, 2004, from http://*www.gao.gov/cgi-bin/getrpt?GAO-04-154*

Verton, D. (2004). Security. *Computerworld, 38*(1), 24-25.

Vijayan, J. (2003). CSOs join forces against cyberthreats. *Computerworld, 37*(46), 12.

Vijayan, J. (2004). New tools help users manage security events. *Computerworld.* Retrieved May 28, 2004, from *http://computerworld.com/securitytopics/security/story/0,10801,90223,00.html*

Weiss, T. R. (2004). GAO hits IT security at USDA, says improvements needed. *Computerworld.* Retrieved May 28, 2004, from *http://computerworld.com/securitytopics/security/story/0,10801,90709,00.html*

Whitman, M. E., & Mattord, H. J. (2003). *Principles of information security.* Boston: Thomson Course Technology.

KEY TERMS

Access: The ability to physically or electronically obtain data or information.

Authentication: Procedures employed to verify the identity of an entity.

Authorization: Privileges afforded to an entity to access equipment and/or information.

Confidentiality: The protection of information from exposure to others.

Data Integrity: The assurance that data has not been tampered with, changing it from its original form.

Firewall: Hardware or software that creates a barrier prohibiting unauthorized access to information systems.

Identity Theft: The acquisition of personal information in order to impersonate the victim in the completion of business or financial transactions.

Information Security: The set of actions taken to protect information from unauthorized access or tampering.

Information System: "[A] discrete set of information resources organized expressly for the collection, processing, maintenance, use, sharing, dissemination, or disposition of information" (Ross et al., 2003).

Network: A "series of interconnected devices and software that allow individuals to share data and computer programs" (U. S. GAO, 2004).

Nonrepudiation: The validation that a message came from a particular source and the inability for the sender to deny responsibility for the message.

Patch: A piece of software code intended to correct a malfunction in an established application.

Risk Assessment: A procedure undertaken to determine the vulnerability of a system.

Threat: Actions that might cause an unauthorized disclosure, modification, or loss of data or information.

ENDNOTE

[1] For the purpose of this article, information is defined as data with meaning.

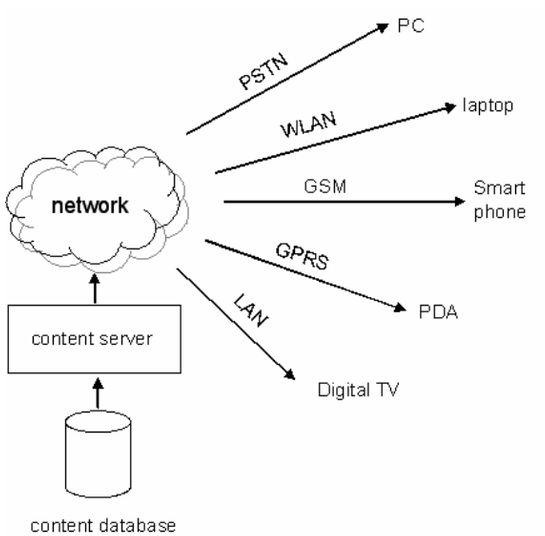
Universal Multimedia Access

Andrea Cavallaro
Queen Mary, University of London, UK

INTRODUCTION

The diffusion of network appliances such as cellular phones, personal digital assistants (PDAs), and handheld computers creates a new challenge for multimedia content delivery: how to adapt the media transmission to various device capabilities, network characteristics, and user preferences. Each device is characterized by certain display capabilities and processing power. Moreover, such appliances are connected through different types of networks with diverse bandwidths. Finally, users with different preferences access the same multimedia content. To cope with the challenge of delivering content to such a variety of conditions while maximizing user satisfaction, multimedia content needs to be adapted to the needs of the specific application, to the capabilities of the connected terminal and network, and to the preferences of the user (Mohan, Smith, & Li, 1999a; Van Beek, Smith, Ebrahimi, Suzuki, & Askelof, 2003). This adaptation enabling seamless access to multimedia content anywhere and anytime is known as universal multimedia access (UMA). The UMA framework is depicted in Figure 1. Three main strategies for adaptive multimedia content delivery have been proposed, namely, the info pyramid, scalable coding, and transcoding. These strategies, emerging trends in UMA and standardization activities, are discussed in the following sections.

Figure 1. Universal multimedia-access framework (PSTN: Public Switched Telephone Network; WLAN: Wireless LAN; LAN: Local Area Network; GSM: Global System for Mobile Communications; GPRS: General Packet Radio Service)



INFO PYRAMID

Traditional solutions to multimedia adaptation encode and store multimedia content in a variety of modalities and formats that are expected to fit possible terminals and networks (Li, Mohan, & Smith, 1998). The most adequate version is then selected for delivery according to the network and hardware characteristics of the specific appliance. The advantage of this approach is speed of access because the content is already available and does not need to undergo any transformations. On the other hand, the limitation of this approach is the difficulty of generating a distinct content version for each profile of capabilities for the large variety of terminals and networks currently available. A general framework for managing and manipulating media objects is the info pyramid. The info pyramid manages different versions, or variations, of media objects with different modalities (e.g., video, image, text, and audio) and fidelities (summarized, compressed, and scaled variations). Moreover, it defines methods for manipulating, translating, transcoding, and generating the content (Smith, Mohan, & Li, 1999b). When a client device requests a multimedia document, the server selects and delivers the most appropriate variation. The selection is based on network characteristics and terminal capabilities, such as display size, frame rate, color depth, and storage capacity. The info pyramid of a media object

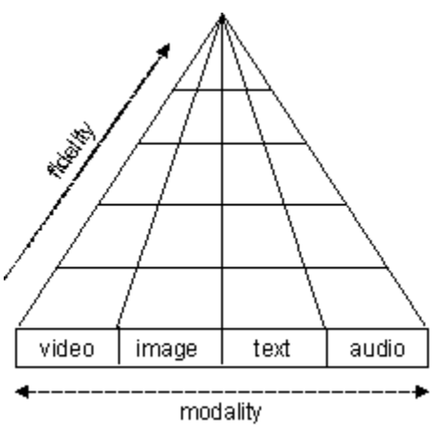
Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

is defined as a collection of the different variations of that media object, as shown in Figure 2. A content value score is then associated to each media object. The value score is assigned manually or based on some automatic measure, such as the entropy. Finally, the most appropriate media object is selected by maximizing the total content value for a set of device and/or network constraints. Utility-based frameworks are generally developed for the selection mechanism. In Mohan, Smith, and Li, (1999), the rate-distortion framework is generalized to a value-resource framework by treating different variations of a content item as different compressions, and different client resources as different bit rates. With the info pyramid approach, higher quality or higher resolution bit streams repeat the information already contained in lower quality or resolution streams. Then additional information is added to manage the streams. For these reasons, the info pyramid is not efficient. To overcome this problem, the redundancy should be removed by coding multiple fidelity levels into a single stream, as described in the next section.

SCALABLE CODING

As opposed to the info pyramid, scalable coding processes multimedia content only once. Lower qualities or lower resolutions of the same content are then obtained by truncating certain layers or bits from the original stream (Wang, Osterman, & Zhang, 2002). In the case of video, basic modes of scalability include quality scalability, spatial scalability, temporal scalability, and frequency scalability. These basic scalability schemes can be combined to reach fine-granularity scalability, such as in MPEG-4 FGS (Fine Granularity Scalability) (Motion Pictures Expert Group; Li, 2001). Quality or SNR (Signal-to-Noise Radio) scalability is defined as the representation of a video sequence with varying accuracies in the color patterns. This is typically obtained by quantizing the color values with increasingly finer quantization step sizes, as shown in Figure 3. Spatial scalability is the representation of the same video in varying spatial resolutions. Corresponding layered bit streams are usually produced by computing a multiresolution decomposition of the original image. Next, the lowest resolution image is coded directly to produce a first layer. For each successive layer, the image from the previous layer is first interpolated to the new resolution, and then the error between the original and the interpolated image is encoded. Temporal scalability is the representation of the same video at varying temporal resolutions or frame rates. The procedure for producing temporally layered bit streams is similar to the procedure used for spatial scalability, but temporal resampling is used instead of spatial resampling. Frequency scalability includes different frequency components in each layer, with the base layer containing low-frequency components and the other layers containing increasingly high-frequency components. Such decomposition can be achieved via frequency transforms like the DCT (Discrete Cosine Transform) or wavelet transforms.

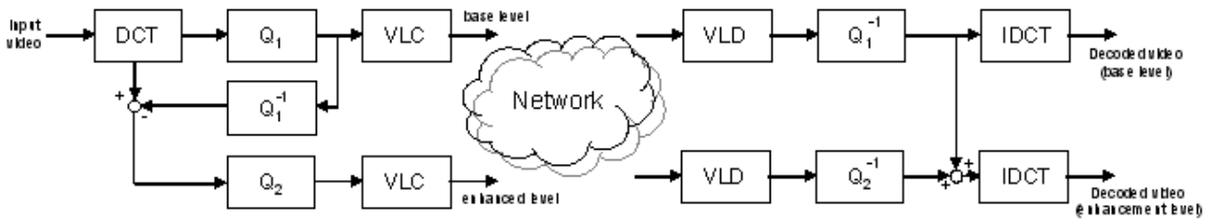
Figure 2. Multimodal representation of a media object as a collection of different variations of the same object in the info-pyramid approach



TRANSCODING

Transcoding is the process of converting a compressed multimedia signal into another compressed signal with different properties (Vetro, Christopoulos, & Sun, 2003). Unlike the info pyramid and scalable coding, transcoding can operate according to the current usage environment on the fly without requiring a priori knowledge of terminal and network capabilities. Early solutions to transcoding determined the output format based on network and appliance constraints only, independent of the se-

Figure 3. Flow diagram of an SNR scalable coder based on two levels. Q_1, Q_2: quantization (Q_2 is finer than Q_1), VLC: variable-length coder (IDCT: inverse discrete cosin transform)

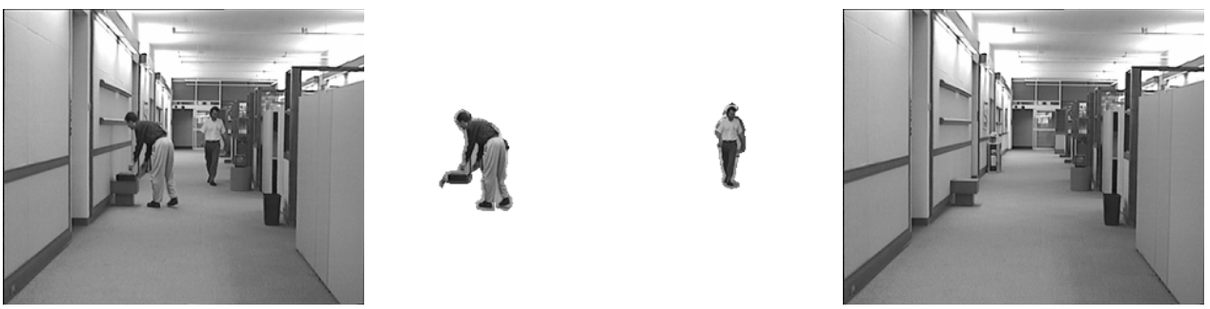


mantics in the content. These solutions are referred to as content-blind transcoding techniques. Content-blind transcoding strategies include spatial resolution reduction, temporal resolution reduction, and bit-rate reduction (Vetro et al.). Spatial resolution reduction affects the size of each frame, thus enabling content distribution to devices with limited display capabilities. Spatial resolution reduction can be obtained by first decoding the video stream and then fully reencoding the reconstructed signal at the new resolution. This approach, referred to as cascaded pixel-domain transcoder, has high memory requirements and is computationally expensive. For this reason, spatial resolution reduction is mostly performed in the frequency domain using motion vector mapping and DCT-domain down-conversion techniques (Shanableh & Ghanbari, 2000; Tan, Liang, & Sun, 2004; Vetro, Hata, Kuwahara, Kalva, & Sekiguchi, 2002). Temporal resolution reduction modifies the frame rate to enable content distribution to devices with limited processing power. Frame-rate reduction is acceptable only when motion activity in the video is limited. When motion activity is high, temporal conversion may limit the impression of motion continuity for the user, thus sensibly reducing the perceived quality. Bit-rate reduction aims at meeting an available channel capacity. As for spatial and temporal resolution reduction, significant complexity savings can be achieved by using simplified frequency-domain architectures. For instance, drift-free MPEG-2-video bit-rate reduction can be performed entirely in the frequency domain by implementing the various modes of motion compensation defined by MPEG-2 in the DCT domain (Assuncao & Ghanbari, 1998). The transcoding strategies described thus far, referred to as intramedia transcoding strategies, do not change the media nature of the input signal. On the other hand, intermedia transcoding (or transmoding) is the process of converting the media input into another media format. Examples of intermedia transcoding include speech-to-text (Morgan & Bourlard, 1995) and video-to-text (Jung, Kim, & Jain, 2004) translation.

SEMANTIC ADAPTATION

The success of UMA applications depends on user satisfaction, which in turn depends on the perceived quality of the content delivered. In order to maximize the perceived video quality, an increasing research effort is aimed at improving coders by taking into account human factors (Lu, Lin, Yang, Ong, & Yao, 2004). The various adaptation methods introduced in the previous sections treat the entire scene uniformly, assuming that people may look at every pixel of the video. In reality, humans do not scan a scene in raster fashion. Our visual attention tends to jump from one point to another. These jumps are called saccades. The saccadic patterns depend on the visual scene as well as on the cognitive task to be performed. For this reason, recent adaptation techniques make use of semantics to minimize the degradation of important image regions (Cavallaro, Steiger, & Ebrahimi, 2003). These techniques attempt to emulate the human visual system to prioritize the visual data in order to improve the performance of the coders. To this end, a scene may be decomposed into objects (Cavallaro, Steiger, & Ebrahimi, 2002) as shown in Figure 4. Then, using object-based temporal scalability (OTS), the frame rate of foreground objects is enhanced so

Figure 4. Automatic decomposition of a scene into background and foreground objects. This decomposition enables semantic adaptation with the separate processing of relevant and less relevant visual information.



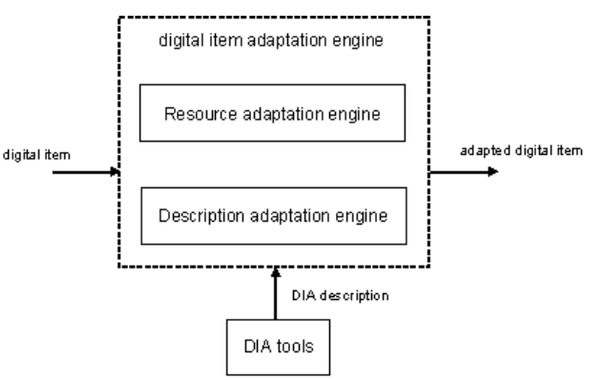
that the foreground has a smoother motion than the background. This is usually achieved by encoding the original video sequence at a low frame rate in a base layer. One or more enhancement layers representing only foreground objects are then encoded so as to achieve a higher frame rate than the base layer. Enhancement frames are coded by predicting from the base layer, followed by overlapping the objects of the enhancement layer on the combined frame. In semantic transcoding, optimal quantization parameters and frame skip can be determined for each video object individually (Vetro, Sun, & Wang, 2001). The bit-rate budget for each object is allocated by a difficulty hint, a weight indicating the relative encoding complexity of each object. Frame skip is controlled by a shape hint, which measures the difference between two consecutive shapes to determine whether an object can be temporally down-sampled without visible composition problems. Key objects are selected based on motion activity and bit complexity. Motion activity and spatial activity descriptors are used as well to combine the requantization of DCT coefficients with spatial down-sampling and temporal down-sampling for content-based, hybrid video transcoding (Liang & Tan, 2001).

An important open issue in semantic adaptation is quality assessment. Perceptual quality assessment is a difficult task already when dealing with traditional coders such as MPEG-1 and MPEG-2 (Olsson, Stroppiana, & Baina, 1997). When dealing with semantic adaptation and user preferences, the task becomes even more challenging. For this reason, a combination of subjective and objective evaluation techniques is usually employed to compare the performance of different adaptation modalities. Traditional PSNR (Personal Signal-to-Noise Ration) analysis uniformly weighs the contribution of each pixel in an image when computing the mean squared error (MSE). Using this analysis, relevant as well as less relevant parts of an image are given the same importance. To account for the way humans perceive visual information, different parts of an image, or object classes, should be considered (Cucchiara, Grana, & Prati, 2002). Object classes are taken into account through a distortion measure, the semantic mean squared error, which assigns a different weight to each semantic class.

STANDARDS

Enabling access to any multimedia content from any type of terminal or network requires the definition and use of standard tools. In order to achieve

Figure 5. MPEG-21 digital item adaptation architecture



interoperable and transparent access to multimedia content, the MPEG standardization committee developed MPEG-21, part 7 (MPEG MDS Group, 2003), which is focused on digital item adaptation (DIA). DIA (Vetro, 2004) aims at providing a set of standardized tools for the adaptation of digital items (Figure 5).

These tools enable the description of the usage environment. This description is based on information about terminal capabilities, network characteristics, user characteristics, and natural environment characteristics. Terminal capabilities include information on device properties and codec capabilities. Device properties are storage and data I/O (input/output) characteristics, and power-related attributes. Codec capabilities specify the format that a terminal is capable of decoding. Network characteristics include network capabilities and network conditions. Network capabilities define the network's static attributes (e.g., minimum guaranteed bandwidth and maximum capacity), whereas network conditions describe dynamic network parameters, such as delay characteristics, available bandwidth, and error. User characteristics include usage history and user preferences, demonstration preference, accessibility characteristics, and location characteristics. Finally, natural environment characteristics pertain to the physical environmental conditions around the user that affect the way the content is consumed. Noise level and lighting conditions are examples of these characteristics. In addition to the above, natural environment characteristics represent time and location. MPEG-21 DIA specifies only the tools that assist with the adaptation process and not the adaptation engine itself, which is left outside the standard to enable the use of new and improved algorithms.

CONCLUSION

In this article, we discussed the concept of multimedia content delivery anywhere and anytime. In particular, we reviewed different forms for implementing UMA, namely, the info pyramid, scalable coding, and transcoding, and we discussed new adaptation forms based on content semantics. The info pyramid and scalable coding operate when the content is prepared. Content preparation aims at matching possible network and terminal capabilities. For this reason, potential profiles of capabilities need to be known a priori. On the other hand, transcoding takes place at the time of delivery. The input bit stream is converted according to the actual needs of the connected appliance and no prior knowledge is required. However, this flexibility comes at the price of a higher computational load. An emerging trend in UMA is the use of semantics that are being introduced in the adaptation mechanism in order to exploit the characteristics of human perception and maximize user experience.

Because accessing multimedia information anywhere and anytime enables an increase in productivity as well as the improvement of user satisfaction through new multimedia services and applications, UMA is having a significant impact on large communities, from corporate to private users.

REFERENCES

Assuncao, P. A. A., & Ghanbari, M. (1998). A frequency-domain video transcoder for dynamic bit-rate reduction of MPEG-2 bit streams. *IEEE Transactions on Circuits and Systems for Video Technology, 8*(8), 953-967.

Cavallaro, A., Steiger, O., & Ebrahimi, T. (2002). Multiple objects tracking in complex scenes. *ACM Multimedia*, 523-532.

Cavallaro, A., Steiger, O., & Ebrahimi, T. (2003). Semantic segmentation and description for video transcoding. *IEEE International Conference on Multimedia and Expo, 3*, 597-600.

Cucchiara, R., Grana, C., & Prati, A. (2002). Semantic transcoding for live video server. *ACM Multimedia*, 223-226.

Jung, K., Kim, K. I., & Jain, A. K. (2004). Text information extraction in images and video: A survey. *Pattern Recognition, 37*(5), 977-997.

Li, C.-S., Mohan, R., & Smith, J. (1998). Multimedia content description in the info pyramid. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 171-178.

Li, W. (2001). Overview of fine granularity scalability in MPEG-4 video standard. *IEEE Transactions on*

Circuits and Systems for Video Technology, 11(3), 301-317.

Liang, Y., & Tan, Y.-P. (2001). A new content-based hybrid video transcoding method. *IEEE International Conference on Image Processing, 1*, 429-432.

Lu, Z., Lin, W., Yang, X. K., Ong, E. P., & Yao, S. S. (2004). Spatial selectivity modulated just-noticeable-distortion profile for video. *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 705-708.

Mohan, R., Smith, J., & Li, C.-S. (1999a). Adapting multimedia Internet content for universal access. *IEEE Transactions on Multimedia, 1*(1), 104-114.

Mohan, R., Smith, J., & Li, C.-S. (1999b). Content adaptation framework: Bringing the Internet to information appliances. *IEEE Global Telecommunications Conference*, 2015-2021.

Morgan, N., & Bourlard, H. (1995). Continuous speech recognition. *IEEE Signal Processing Magazine, 12*(3), 24-42.

MPEG MDS Group. (2003). *MPEG-21 multimedia framework, part 7: Digital item adaptation* (ISO/MPEG N5845). Retrieved July 2, 2004, from *http://www.chiariglione.org/mpeg/working_documents/mpeg-21/dia/dia_fcd.zip*

Olsson, S., Stroppiana, M., & Baina, J. (1997). Objective methods for assessment of video quality: State of the art. *IEEE Transactions on Broadcasting, 43*(4), 487-495.

Shanableh, T., & Ghanbari, M. (2000). Heterogeneous video transcoding to lower spatio-temporal resolutions and different encoding formats. *IEEE Transactions on Multimedia, 2*(2), 101-110.

Smith, J., Mohan, R., & Li, C.-S. (1999). Scalable multimedia delivery for pervasive computing. *ACM Multimedia, 1*, 131-140.

Tan, Y.-P., Liang, Y., & Sun, H. (2004). On the methods and performances of rational downsizing video transcoding. *Signal Processing: Image Communications, 19*, 47-65.

Van Beek, P., Smith, J., Ebrahimi, T., Suzuki, T., & Askelof, J. (2003). Metadata-driven multimedia access. *IEEE Signal Processing Magazine, 48*(3), 40-52.

Vetro, A. (2004). MPEG-21 digital item adaptation: Enabling universal multimedia access. *IEEE Multimedia*, 84-87.

Vetro, A., Christopoulos, C., & Sun, H. (2003). Video transcoding architectures and techniques: An overview. *IEEE Signal Processing Magazine, 20*(2), 18-29.

Vetro, A., Hata, T., Kuwahara, N., Kalva, N., & Sekiguchi, S.-I. (2002). Complexity-quality analysis of transcoding architectures for reduced spatial resolution. *IEEE Transactions on Consumer Electronics*, 515-521.

Vetro, A., Sun, A., & Wang, Y. (2001). Object-based transcoding for adaptable video content delivery. *IEEE Transactions on Circuits and Systems for Video Technology, 11*(3), 387-401.

Wang, Y., Ostermann, J., & Zhang, Y.-Q. (2002). *Video processing and communications* (1st ed.). Prentice Hall.

KEY TERMS

Info Pyramid: Multimedia data representation based on storing different versions of media objects with different modalities and fidelities.

Intermedia Transcoding: The process of converting the media input into another media format.

Intramedia Transcoding: A transcoding process that does not change the media nature of the input signal.

MPEG: Motion Pictures Expert Group.

MPEG-1: Standard for the coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s. MPEG-1 is the standard on which video CD and MP3 are based.

MPEG-2: Standard for the generic coding of moving pictures and associated audio information. MPEG-2 is the standard on which digital television set-top boxes and DVD are based.

MPEG-4: Standard for multimedia for fixed and mobile Web.

MPEG-7: Multimedia content-description interface. MPEG-7 is the standard for the description and search of audio and visual content.

MPEG-21: Multimedia framework initiative that enables the transparent and augmented use of multimedia resources across a wide range of networks and devices.

Transcoding: The process of converting a compressed multimedia signal into another compressed signal with different properties.

UMA: Universal multimedia access.

Usability

Shawren Singh
University of South Africa, South Africa

INTRODUCTION

In this article we will examine some important issues related to human-computer interaction (HCI). This will be followed by a discussion of usability and its underlying principles and properties. The dependability of computer systems is intrinsically multi-faceted. Dependable hardware is patently of limited value unless accompanied by dependable software. Neither helps greatly if human interaction with the hardware and software system is fault-prone and the dependable socio-technical performance of an inappropriate task may cause wider damage (MacKenzie, 2000).

HCI

HCI is a field of research and development, methodology, theory and practice, with the objective of designing, constructing and evaluating computer-based interactive systems – including hardware, software, input/output devices, displays, training and documentation – so that people can use them efficiently, effectively, safely and with satisfaction. HCI is cross-disciplinary in its conduct and multidisciplinary in its roots, drawing on – synthesizing and adapting from – several other fields, including human factors (e.g., the roots of task analysis and designing for human error in HCI), ergonomics (e.g., the roots of design of devices, workstations and work environments), cognitive psychology (e.g., the roots of user modeling), behavioral psychology and psychometrics (e.g., the roots of user performance metrics), systems engineering (e.g., the roots of much pre-design analysis) and computer science (e.g., the roots of graphical interfaces, software tools and issues of software architecture)(Hartson, 1998).

HCI is a simple concept, but difficult to explain in just a few sentences. This difficulty has to do in part with its origins – it draws upon many different but related disciplines that make it a true multi- and inter-disciplinary field. However, as the name indicates, it is a field that concerns itself primarily with what happens when humans and computers meet.

USABILITY

The success of any interactive product or system is ultimately dependent on it providing the right facilities for the task at hand in such a way that it can be effectively used, at an a appropriate price (Dillon, 1994). In the past, implementation of business software involved acquiring (purchasing) a piece of software in order to install and use an interface. The World Wide Web (WWW) and its associated standardized technologies have changed this and now provide the interface before transactions are deployed (Singh & Kotze, 2002).

Usability is generally regarded as ensuring that interactive products, such as e-commerce applications, are easy to learn, effective to use, have aesthetic integrity, are enjoyable from the user's perspective and involve the optimization of user interaction with these interactive products (Preece, Rogers & Sharp, 2002). Over time, several researchers have produced sets of generic usability principles that can be used in improving e-commerce Web sites as well as showing how to test usability and how to design software products, bearing usability in mind (Badre, 2002; Cato, 2001; Dix, Finlay, Abowd & Beale, 1998; Mayhew, 1999; Nielsen, 1993, 2000; Preece et al., 2002; Preece, Rogers, Sharp, Benyon, Holland & Carey, 1994; Shneiderman, 1998; Thimbley, 1990). These principles include aspects such as effectiveness, efficiency, safety, utility, learnability, flexibility, robustness, memorability and so forth.

ISO Definition

The ISO 9241 (ISO 9241, 1998; Travis, 2003) standard describes ergonomic requirements for office work with visual display terminals. This standard

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

defines how to specify and measure the usability of products, as well as defining the factors that have an effect on usability.

In order to specify or measure usability, it is necessary to identify the goals that pertain to and decompose effectiveness, efficiency and satisfaction, and the components of the context of use into subcomponents with measurable and verifiable attributes:

- **Effectiveness:** the accuracy and completeness with which specified users can achieve specified goals in particular environments
- **Efficiency:** the resources expended in relation to the accuracy and completeness of goals achieved
- **Satisfaction:** the comfort and acceptability of the work system to its users and other people affected by its use.

The standard states that when specifying or measuring usability, the following information is needed: a) A description of the intended goals; b) a description of the components of the context of use, including users, tasks, equipment and environments (This may be a description of an existing context, or a specification of intended contexts. The relevant aspects of the context and the level of detail required will depend on the scope of the issues being addressed. The description of the context needs to be sufficiently detailed so that those aspects of the context that may have a significant influence on usability can be reproduced.); and c) Target or actual values of effectiveness, efficiency and satisfaction for the intended contexts.

The context of use defined by the standard includes the following factors:

- **Description of users:** Characteristics of the users need to be described. These can include knowledge, skill, experience, education, training, physical attributes, and motor and sensory capabilities. It may be necessary to define the characteristics of different types of users; for example, users having different levels of experience or performing different roles.
- **Description of tasks:** Tasks are the activities undertaken to achieve a goal. Characteristics of tasks that may influence usability should be described; for example, the frequency and duration of the task. Detailed descriptions of the activities and processes may be required if the description of the context is to be used as a basis for the design or evaluation of details of interaction with the product. This may include descriptions of the allocation of activities and steps between the human and technological resources. Tasks should not be described solely in terms of the functions or features provided by a product or system. Any description of the activities and steps involved in performing the task should be related to the goals that are to be achieved.
- **Description of equipment:** The description of the hardware, software and materials may take place in terms of a set of products, one or more of which may be the focus of usability specifications or evaluation, or it may occur in terms of a set of attributes or performance characteristics of the hardware, software and other materials.
- **Description of environment:** Relevant characteristics of the physical and social environment need to be described. Aspects that may need to be described include attributes of the wider technical environment (e.g., the local area network), the physical environment (e.g., workplace, furniture), the ambient environment (e.g., temperature, humidity) and the social and cultural environment (e.g., work practices, organizational structure and attitudes).
- **Usability measures:** Usability measures include effectiveness, efficiency and satisfaction. These are measured in user trials of the product. The goal of the user trial is to identify what the projects are trying to find out in their trials. The goal of the user trial may be to help in defining user requirements, to confirm that the technology works in real conditions, to measure user attitudes, to start the marketing of the system or to measure accessibility.

USABILITY IN GENERAL

Guidelines are lists of rules about when and where to do things, or not to do things, in an interface. These guidelines can take a variety of forms and may be obtained from several sources, such as journal articles, general textbooks, company in-house style guides and so forth. (Singh, Erwin & Kotze, 2001).

Dix et al. (1998), for example, put forward principles to support usability in three categories: *Learnability*, *flexibility* and robustness. *Learnability* refers to the ease with which new users can begin effective interaction and then attain a maximal level of performance. Usability principles related to learnability include predictability, synthesizability, familiarity, generalisability and consistency. *Flexibility* refers to the multiplicity of ways in which the user and the system exchange information. A user is engaged with a computer to achieve some set of goals in the work or task domain. Usability principles related to flexibility include dialogue initiative, multithreading, task migratability, substitutivity and customisability. *Robustness* refers to the level of support given to the user in determining successful achievement and assessment of goals. Usability principles related to robustness include observability, recoverability, responsiveness and task conformance.

Shneiderman (1998) also focused on this aspect. He advocates three groups of principles when he discusses user-centered design. Many of these overlap with the principles proposed by Dix et al. (1998). Shneiderman's principles include recognition of diversity, use of the eight golden rules of interface design (see the following) and prevention of errors.

The eight golden rules for interface design (Shneiderman, 1998) constitute the underlying principles of design applicable to most interactive systems. These underlying principles must be interpreted, refined and extended for each environment and include striving after consistency; enabling frequent users to use shortcuts; offering informative feedback; designing dialogues to yield closure (the completion of a group of actions); offering error prevention and simple error handling; permitting easy reversal of actions; supporting internal locus of control; and reducing short-term memory load.

All applications require user interfaces, the design of which is not a trivial matter. The same is true for e-commerce and any other Web-based applications. Shneiderman (1998) states that within the ocean (WWW) of information, "there are also lifeboat Web pages offering design principles, but often the style parallels the early user-interface writings of the 1970s." The problem of early user interfaces, ignoring the abilities and preferences of the users, is therefore still present.

WEB USABILITY

Usability has assumed a much greater importance in the Internet economy than it has in the past. In traditional physical product development, customers did not get to experience the usability of the product until after they had already bought and paid for that product. Usability rules the Web. Simply stated, if customers cannot find their desired product, they will not buy it. The Web is the ultimate customer-empowering environment. The controller of the mouse gets to decide everything. It is so easy to go elsewhere – all the competitors in the world are but a mouse-click away (Nielsen, 2000).

Nielsen (2000) identifies the following common errors made in Web design:

- **Business models:** treating the Web as a Marcom brochure instead of a fundamental shift that will change the way we conduct business in the network economy.
- **Project management:** managing a Web project as if it were a traditional corporate project. This leads to an internally focused design with an inconsistent user interface. Instead, a Web site should be managed as a single customer-interface project.
- **Information architecture:** structuring the site to mirror the way the company is structured instead of mirroring the users' tasks and their views of the information space.
- **Page design:** creating pages that look gorgeous and that evoke positive feelings when demonstrated inside the company. Internal demos do not suffer the response-time delays that are the main determinant of Web usability. Similarly, a demo does not expose the difficulties a novice user will have in finding and understanding the various page elements. Instead, design for an optimal user experience under realistic circumstances, even if your demos will be less "cool."
- **Content authoring:** writing in the same linear style as you have always written. Instead, force yourself to write in the new style that is optimized for online readers who frequently scan text and who need very short pages with secondary information relegated to supporting pages.

- **Linking strategy:** treating your own site as the only one the matters, without proper links to other sites and without well-designed entry-points for others to link to. Many companies do not even use proper links when they mention their own site in their own advertising. Instead, remember that hypertext is the foundation of the Web and that no site is an island.

Usability Methods for the Web

There are at least five ways to incorporate usability testing in the Web development process: *heuristic evaluation, usability tests, user and task analysis, checklists* and *walkthroughs.*

- *Heuristic evaluation* is sometimes referred to as "discount usability engineering." It is a form of usability inspection in which usability specialists evaluate whether an interface follows established usability guidelines, or heuristics. Heuristic evaluation is best conducted in early, prototype stages of an e-commerce project and repeated as significant design changes are implemented.
- *Usability tests* are true tests when conducted with actual end users, ideally in their own environment, while performing real tasks. Only three to five users are needed to obtain a significant amount of valuable data. Research conducted by Nielsen (1993) has shown that as many as 85% of usability problems can be identified in the first usability test with a small group. Once the site has been redesigned based on the initial usability test results, it must be tested again to ensure that prior problems have been ironed out and new problems have not been introduced.
- *User and task analysis* is the process of learning about ordinary users by observing them in action. User and task analysis is conducted before design begins. The results of the analysis will be used to create the information architecture, navigation structure and labeling schemes that make sense to users.
- *Checklists* constitute a subjective evaluation approach, closely related to the heuristic evaluation methods as examined in the preceding section. This approach is based on training, field experience and an examination of human factors data. The above Web usability testing techniques will be further discussed (van Dyk, 1999)
- The goal of *Walkthroughs* in HCI design is to detect problems very early on so they can be eliminated (Preece et al., 1994).

CONCLUSION

From the user's perspective, usability is important because it can make the difference between performing a task accurately and completely or not, and enjoying the process or being frustrated. From the developer's perspective, usability is important because it can mean the difference between the success and failure of a system. From a management point of view, software with poor usability can reduce the productivity of the workforce to a level of performance worse than it would be without the system. In all cases, lack of usability can cost time and effort, and can largely determine the success or failure of a system. Given a choice, people will tend to buy systems that are more user friendly (W1, 2004).

REFERENCES

Badre, N.A. (2002). *Shaping Web usability: Interaction design in context*. Boston: Addison-Wesley.

Cato, J. (2001). *User-centered Web design*. Harlow: Addison-Wesley.

Dillon, A. (1994). *Designing usable electronic text: Ergonomic aspects of human information usage*. London: Taylor & Francis Ltd.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (1998). *Human-computer interaction* (2nd ed.). Harlow, England: Prentice Hall.

Hartson, R.H. (1998). Human-computer interaction: Interdisciplinary roots and trends. *The Journal of Systems and Software, 43*, 103-118.

ISO 9241. (1998). Ergonomic requirements for office work with visual display terminals: The International Organization for Standardization.

MacKenzie, D. (2000). A view from the Sonnenbichl: On the historical sociology of software and system dependability. Paper presented at the *International Conference on the History of Computing: Software Issues*, Paderborn, Germany.

Mayhew, D.J. (1999). *The usability engineering lifecycle: a practitioner's handbook for user interface design*. San Francisco: Morgan Kaufmann.

Nielsen, J. (1993). *Usability engineering*. San Diego: Morgan Kaufmann.

Nielsen, J. (2000). *Designing Web usability: The practice of simplicity*. Indianapolis: New Riders.

Preece, J., Rogers, Y., & Sharp, H. (2002). *Interaction design: Beyond human-computer interaction*. New York: John Wiley & Sons.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., & Carey, T. (1994). *Human-computer interaction*. Harlow, England: Addison-Wesley.

Shneiderman, B. (1998). *Design the user interface: Strategies for effective human-computer interaction* (3rd ed.). Reading: Addison-Wesley.

Singh, S., Erwin, G., & Kotze, P. (2001). Electronic Business Accepted Practices (e-BAP): Standardised HCI for e-commerce in South Africa. Paper presented at the *Postgraduate Research Symposium: SAICSIT 2001*, Pretoria.

Singh, S., & Kotze, P. (2002, September 16-18). Towards a framework for e-commerce usability. Paper presented at the *SAICSIT: Enablement Through Technology*, Port Elizabeth, South Africa.

Thimbley, H. (1990). *User interface design*. Wokingham, England: Addison-Wesley.

Travis, D. (2003). Bluffers' guide to ISO 9241. Userfocus Ltd. Retrieved August 19, 2003, from *www.userfocus.co.uk*

van Dyk, T. (1999). *Usability and Internet-based banking.* Unpublished masters. University of South Africa, Pretoria.

W1. (2004). Usability first. Diamond Bullet. Retrieved June 29, 2004, from *www.usabilityfirst.com/intro/index.txl*

KEY TERMS

Accessibility: Just as computers vary by operating system, processor speed, screen size, memory and networking abilities, users vary in ways both expected and unexpected. Some differences more commonly thought of are language, gender, age, cultures, preferences and interests. However, some of the differences that need to be paid more attention by the software and Web development community are skills, ability levels and constraints under which users may be operating. Designing for diversity not only increases the number of people able to access software or a Web site but also increases their level of involvement with it.

Aesthetic Integrity: A principle that advocates that a design should be visually appealing and should follow common principles of visual design – consistency, a clear identity, a clear visual hierarchy, good alignment, contrast and proportions.

E-Commerce: Uses some form of transmission medium through which exchange of information takes place in order to conduct business.

HCI: A field of research and development, methodology, theory and practice with the objective of designing, constructing and evaluating computer-based interactive systems – including hardware, software, input/output devices, displays, training and documentation – so that people can use them efficiently, effectively, safely and with satisfaction.

Interactive Systems: These support communication in both directions, from user to computer and back. A crucial property of any interactive system is its support for human activity.

ISO 9241: This standard describes ergonomic requirements for office work with visual display terminals. It defines how to specify and measure the usability of products, and defines the factors that have an effect on usability.

Usability: Generally regarded as ensuring that interactive products, such as e-commerce applications, are easy to learn, effective to use and enjoyable from the user's perspective. It involves the

optimization of user interaction with these interactive products.

User-Centered Design: The real users and their goals, not just technology, should be the driving force behind the development of a product.

Usability Assessment in Mobile Computing and Commerce

Kuanchin Chen
Western Michigan University, USA

Hy Sockel
Youngstown State University, USA

Louis K. Falk
Youngstown State University, USA

USABILITY STANDARDIZATION

Usability is an acknowledged important aspect of any system or product design. Researchers have found that a good interface design promotes higher mutuality (feeling similar and connected), which, in turn, leads to higher levels of involvement and a favorable impression of credibility.

Many practitioners and researchers (Nielsen, 2000) have elaborated on usability aspects, but few have agreed upon a unifying definition. In 1998 the International Organization for Standardization (ISO) defined usability as follows:

Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. (ISO 9241-11, 1998, p.2)

From this definition, it can be construed that effectiveness, efficiency, and satisfaction are three pillars for usability measures. In this regard, the ISO defines:

- effectiveness as the "accuracy and completeness with which users achieve specified goals";
- efficiency as the "resources expended in relation to the accuracy and completeness with which users achieve goals"; and
- satisfaction as the "freedom from discomfort, and positive attitudes towards the use of the product."

The ISO standard acknowledges that the level of usability depends highly on the intended context of use (e.g., users, hardware, software, and social environments). Researchers have demonstrated that the three ISO usability components are distinct. Frøkjær et al. (2000) found only a weak relationship among the three usability components. Walker et al. (1998) found that efficiency did not translate into user satisfaction. These empirical studies suggest that efficiency, effectiveness, and satisfaction may be independent aspects of usability, and a causal relationship among them may be weak or even nonexistent.

OTHER DIMENSIONS OF USABILITY

Research has not been limited to the three main ISO characteristics. Researchers such as Sing (2004), Hilbert and Redmiles (2000), and McLaughlin and Skinner (2000) support ISO standard's recommendation that usability is highly contextual and built on factors such as the users' past experiences with similar systems, the role they play, and the environment in which the product is used. In addition, users' expectations and priorities toward usability also depend on the role they play and the position they hold.

Sing (2004) cites studies that include software usability components of (a) flexibility (users perceive that the system can adapt to their preferred style of interaction); (b) easy to learn (users perceive that it is easy to gain required knowledge to achieve a satisfactory level of competence); and (c)

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

easy to remember (it is easy for users to recall system features after a period of time).

Hilbert and Redmiles (2000) offer similar dimensions of usability: (a) learnability (the system is easy to learn); (b) efficiency (the system is efficient to use); once a user masters the system; a higher level of productivity is possible; (c) memorability (the system should be easy to remember, even for casual users); (d) errors (the system should have a low error rate); and (e) satisfaction (the system should be pleasant to use).

McLaughlin and Skinner (2000) examined six usability components on new IT implementations: (a) checkability (the system's ability to ensure information correctness); (b) confidence (the users' confidence in their ability to use the system and in the system itself); (c) control (system offers the users control); (d) ease of use; (e) speed of use; and (f) understanding.

USABILITY EVALUATION METHODS AND INSTRUMENTS

Evaluation Methods

The approach undertaken for usability varies, depending on the intended goals. Ivory and Hearst (2001) outlined a taxonomy view of usability test methods as follows:

- **Method Class:** Testing, inspection, inquiry, analytical modeling, and simulation.
- **Method Type:** Log file analysis, guideline review, surveys, GOMS analysis, genetic algorithms, and so forth.
- **Automation Type:** None, capture, analysis, critique.
- **Effort Level:** Minimal effort, model development, informal use, and formal use.

Table 1. Usability Instruments

Instrument	Application	Usability Dimension
(1) Software usability measurement inventory (SUMI) (Kirakowski & Corbett, 1993)		
SUMI is intended as an instrument to measure perceived software quality from the end-user standpoint. SUMI consists of 50 questions measuring quality of use in five usability aspects.	Software	Efficiency, effect, helpfulness, control, and learnability.
(2) Web site analysis and measurement inventory (WAMMI) (http://www.wammi.com)		
WAMMI consists of 20 questions to measure the five aspects of Web site usability. The assessment result is compared to a database of similar Web sites to generate the final overall usability rating.	Web sites	Attractiveness, controllability, efficiency, helpfulness, and learnability.
(3) Measuring the usability of multi-media systems (MUMMS) (http://www.ucc.ie/hfrg/questionnaires/mumms/index.html)		
MUMMS targets the assessment of use quality in multimedia systems. It uses the same usability dimensions as SUMI.	Multimedia systems	Efficiency, effect, helpfulness, control, and learnability.
(4) Usability task questionnaire (Sing, 2004)		
Sing's usability task questionnaire consists of 25 Likert-type questions and two open-ended questions. The goal of this questionnaire is to assess six usability components.	Electronic stores	Effectiveness, efficiency, flexibility, easy to learn, easy to remember, satisfaction
(5) WebQual (Barnes &Vidgen, 2002; Barnes, Liu & Vidgen, 2001)		
WebQual is an instrument based on quality function deployment (QFD), which is a structured process to capture "voice of the customer" through each state of product or service development. The current version of WebQual is a 23-question instrument to measure the three quality dimensions of Web sites.	WAP and Web sites	Information quality, interaction and service quality, and usability

Interested readers should consult Ivory and Hearst's study for more details.

Usability Instruments

As with most assessment procedures, usability assessments depend highly on how closely the instrument follows or achieves the intended goals. Since there is a strong tie between the context of use, usability goals, and the measuring instruments, it is difficult to build a comprehensive usability instrument for all circumstances. This section explores a brief survey of usability instruments with a focus on software usability.

USABILITY IN MOBILE COMPUTING

Mobile wireless devices enabled by cell-phone technology, Portable Computing Devices (PCD), Personal Digital Assistance (PDA), Global Position Satellites (GPS), and Geographic Information Systems (GIS) are being used online to create a mobile commerce (m-commerce) environment. While features that these devices support bear a high level of resemblance to their wired cousins, there are important and significant differences (i.e., wireless devices are ubiquitous in nature and support pervasive computing). However, these devices are limited in the quality and the type of applications they can perform because of reduced communication bandwidth available to them and the physical characteristics of the devices. Furthermore, these mobile devices often are constrained by limited computing capacity, display areas, data entry capability, and power.

The emergence of these mobile devices has pushed vendors, users, and system designers to rethink how the interfaces are put to their best use. For example, cell phones' limited screen display does not allow for extensive animation or color choices. Although modern PDAs offer a larger screen display and better computing power, they are not suited for intensive computing tasks. Not all applications are suited for mobile use; it would be difficult at this time to try mobile devices for activities that demand intense concentration with very accurate manual input (Golenko & Merrick, 2003).

Lee and Benbasat (2004) indicate that "human-computer interaction (HCI) researchers have explored interface designs for mobile devices through which users experience a very different environment than with personal computers." (p. 79). While there may not be enough research to support a set of usability standards for m-commerce, there is large agreement that behavior and devices are sufficiently different and that the design interface standards of e-commerce should not be blindly applied to m-commerce (Brewster, 1998).

Lee and Benbasat (2004) addressed m-commerce usability issues using Rayport and Jaworski's (2001) "7C" framework. The 7C framework was developed for analyzing e-commerce interfaces and examines the customer interface based on seven factors:

1. **Context:** Captures how the Web site is delivered.
2. **Content:** Focuses on what a site delivers.
3. **Community:** Concerns the interaction between users and includes the feeling of membership and a sense of involvement.
4. **Customization:** Refers to the site's ability for personalization, either by allowing the user to tailor the site or to tailor the site to a user (by profiling).
5. **Communication:** Defined as the dialogue between sites and their users.
6. **Connection:** Refers to the extent of formal linkage from one site to others.
7. **Commerce:** Refers to the interface that supports business transactions.

However, Lee and BenBasat (2004) found that a considerable amount of issues in m-commerce is not covered by the 7C's framework. They identified two additional elements they labeled as the "2Ms": (1) mobile setting and (2) mobile device constraints.

Mobile commerce happens in an environment where users are able to do computing/commerce from nearly anywhere and whenever they want with a large variety of computing devices (Lyytinen & Yoo, 2002). The nomadic ubiquitous nature of mobile devices (i.e., phones, PDAs, handhelds, wearable computing devices, and wristwatches)

has limited the assumptions that developers can make about the characteristics of both the users and their devices. Consequently, m-commerce requires fundamentally different approaches in analysis and design that focus not on product itself, but on the use of a product (Golenko & Merrick, 2003). The following is a summarization of some of the usability issues concerning mobile services.

Device Graphics Capabilities

Although a picture can be worth a thousand words, it is not quite true in mobile devices. Graphic capabilities are not only constrained by computing power of mobile devices, but also by display size, network speed, and color support. Ramsay's (2001) study of Wireless Application Protocol (WAP) sites uncovers some usability problems for graphics (graphics used as a splash screen may take too long to load). Graphical headers or logos may take up too much screen real estate, adding little to enhance surfing experience. They also leave little room for text content. To leverage the limitation of graphic display in mobile devices and the need for graphics presentations in certain software applications, Rist and Brandmeier (2002) experimented with eight types of transformations to make low-quality (but comprehensible) graphics for mobile devices. These low-quality graphics are smaller in size than their full-blown originals. The best transformation scheme without degrading much in human comprehension includes the following sequence of graphic operations: scaling of the graphics, color-reduction, and WBMP conversion.

Visibility and Predictability

Visibility of important features and predictability of link destination as well as clear indication of the length of long lists are among the 10 mobile usability principles proposed in Condos et al. (2002). Although scrolling is less favorable on regular home pages (Falk, Sockel, &Chen, in press), scrolling on a long list of options in mobile devices is even more troublesome. Jones et al. (1999) found the size of a display screen had a significant impact on user performance. The smaller the screen, the more the user needed to scroll, and the lower the performance.

Designers of Web (or WAP) content tend to divide a large chunk of information into smaller pages in order to increase usability and, at the same time, to prevent cluttering. Predictability has to do with how these smaller pages are linked together and how the users are provided with some cues to help predict the number of remaining pages in the series. A simplistic approach is to add "bread crumbs," or a small indicator, on the bottom of the page to show where the user is in the series of Web pages. However, page indicators such as "page 3 of 10" do not help when a page requires horizontal scrolling.

Appropriate Content Delivery and Navigation

Content is regarded as the most important aspect of online systems. Without an appropriate navigation design, user disorientation and frustration may likely happen. The WAP environment for mobile devices is no exception. Condos et al. recommend the following usability principles for WAP services: avoid dead links, provide clear and meaningful error messages; present content appropriately and consistently; and deliver content to aid user interaction and to minimize user inputs. Buchnan et al. (2001) studied three types of menu displays for small screens: (a) horizontal scrolling to cycle through the text of menu items; (b) vertical scrolling to drop down the text of menu items; and (c) page scrolling to deliver fewer menu items but more text for each menu item. They found that vertical scrolling was the best among the three in terms of time spent on information access and the error rate.

Experience Enhancement

Venkatesh, Ramesh, and Massey (2003) suggest that content is still "king" in both Web and WAP environments, but relevance, structure, and personalization are critical for better use experience. They further indicated that usability testers and system designers should pay special attention to the reasons (i.e., time pressure, location, and convenience) why mobile devices are used. Lack of screen space can be overcome in part with sound feedbacks to reflect the current state of user interaction. Because mobile users are frequently involved in multi-tasking, some

researchers suggest the use of audio feedback (Brewster, Leplatre & Crease, 1998). This approach may have an additional benefit in that it can help reduce display clutter, allowing for the presentation of more information (Walker & Brewster, 1999). Brewster et al. (1998) recommend the adoption of language-independent non-speech tones (e.g., a beeping sound). Brewster (2002) found that the button size in PDA systems can be reduced from 16 x 16 pixels to 8 x 8 pixels without loss of user performance, if they are sonically-enhanced buttons. However, the use of sound cannot offset the degradation of usability if the button size is further reduced to 4x4 pixels.

As the ISO usability standard suggests, context of use is an integral part that should be taken into account when conducting usability studies. However, analysis of use contexts for the mobile environment is significantly more complex than that of non-mobile environments. Gorlenko and Merrick (2003) outlined the three challenges that face designers of mobile applications and the usability testers.

- The challenge of identifying all possible usages of the mobile products. The rationale is that the more prevalent and convenient the mobile devices are in a particular setting, the more likely the users will be to try to use the devices in a different setting as well.
- The challenge of the changing nature of task environments. The environment where a task is performed may change over the course of the task. Other factors that change very likely may affect the user's task. For example, the weather condition, the network connection, the bandwidth availability, and the number and types of applications running on a multi-tasking system may all cause variations of usability results.
- The challenge of human multi-tasking nature in mobile interactions. An integral assumption of mobile devices is that the user interacts with the device while simultaneously undertaking other tasks. Distractions are the nature of use, rather than an exception. Usability testing also should carefully consider the variety of parallel activities.

CONCLUSION

It is the changing nature of businesses with mobile considerations that are becoming the frontiers. Mobile devices permit shopping, online banking, stock trading, and even gambling from anywhere that wireless is supported. The ubiquitous nature of the devices presents opportunities such as targeted marketing, personalized content delivery, and mass customization. We end by reminding everyone that "for a mobile solution to be successful everyone involved in the development of various components must focus on the total user experience in general, and on usability in particular" (Gorlenko & Merrick, 2003, p. 640).

REFERENCES

Barnes, S.J., Liu, K., & Vidgen, R.T. (2001). Evaluating WAP news sites: The WebQual/M approach. *Proceedings of the Ninth European Conference on Information Systems*, Bled, SLovenia.

Barnes, S., & Vidgen, R. (2002). An integrative approach to the assessment of e-commerce quality, *Journal of Electronic Commerce Research, 3*(3), 114-127.

Brewster, S. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing*, *6*(3), 188-205.

Brewster, S., Leplatre, G., & Crease, M. (1998); Using non-speech sounds in mobile computing devices. *Proceedings of the First Workshop on Human Computer Interaction with Mobile Devices*, Glasgow.

Buchnan, G., Farrant, S., Marsden, G., & Pazzani, M. (2001). Improving mobile Internet usability. *Proceedings of the Tenth International Conference on World Wide Web*, Hong Kong.

Condos, C., James, A., Every, P., & Simpson, T. (2002). Ten usability principles for the development of effective WAP and m-commerce services. *Aslib Proceedings*, *54*(6), 345-355.

Falk, L.K., Sockel, H., & Chen, K. (2004). E-commerce and consumer's expectations. *Journal of Website Promotion*, in press.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 345-352), The Hague, The Netherlands.

Golenko, L., & Merrick, R. (2003). No wires attached: Usability challenges in the connected mobile world. *IBM Systems Journal*, *42*(4), 639-651.

Hilbert, D.M., & Redmiles, D.F. (2000). Extracting usability information from user interface events. *ACM Computing Surveys*, *32*(4), 384-421.

ISO 9241-11. (1998). Ergonomic requirements for office work with visual display terminals (VDTs)—Part 11: Guidance on usability. *ISO/IEC 9241-14: 1998 (e)*. Available at *http://iso.gov*

Ivory, M.Y., & Hearst, M.A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, *33*(4), 470-516.

Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K., & Buchanan, G. (1999). Improving Web interaction on small displays. *Proceedings of the Eighth International Conference on World Wide Web.* Amsterdam, Holland.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, *24*(3), 210-214.

Lee, Y.E., & Benbasat, I. (2004). A framework for the study of customer interface design for mobile commerce. *International Journal of Electronic Commerce*, *8*(3), 79-102.

Lyytinen, K, & Yoo, Y. (2002). Research commentary: The next wave of nomadic computing. *Information Systems Research*, *13*(4), 377-388.

McLaughlin, J., & Skinner, D. (2000). Developing usability and utility: A comparative study of the users of new IT. *Technology Analysis & Strategic Management*, *12*(3), 413-423.

Nielsen, J. (2000). *Designing Web usability*., Indianapolis, IN: New Riders Publishing.

Ramsay, M. (2001). Mildly irritating: A WAP usability study. *Aslib Proceedings*, *53*(4), 141-158.

Rayport, J., & Jaworski, B. (2001). *Introduction to e-commerce.* New York: McGraw-Hill.

Rist, T., & Brandmeier, P. (2002). Customizing graphics for tiny displays of mobile devices. *Personal and Ubiquitous Computing*, *6*, 260-268.

Sing, C.-K. (2004). The measurement, analysis, and application of the perceived usability of electronic stores. *Singapore Management Review*, *26*(2), 49-64.

Venkatesh, V., Ramesh, V., & Messey, A. (2003). Understanding usability in mobile commerce. *Communications of the ACM*, *46*(12), 53-56.

Walker, A., & Brewster, S. (1999). Extending the auditory display space in handheld computing devices. *Proceedings of the Second Workshop on Human Computer Interaction with Mobile Devices*, Edinburgh.

Walker, M.A., Fromer, J., Di Fabbrizio, G., Mestel, C., & Hindle, D. (1998). What can I say? Evaluating a spoken language interface to email. *Proceedings of CHI 98*, Los Angeles, California.

KEY TERMS

Context of Use: In ISO 9241 definition of usability, the three usability components (i.e., effectiveness, efficiency, and satisfaction) are to be examined in a usability study against the context of use. According to the ISO standard 9241-11, context of use refers to "users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a product is used" (p. 2).

GOMS: GOMS is a set of techniques for modeling human task performance. It stands for Goals, Operators, Methods, and Selection rules. GOMS was invented by Card, Moran, and Newell (1983) and has become an important model for Human-Computer Interaction studies.

ISO: International Organization for Standardization (ISO) is a non-governmental organization consisting of standards institutes of 148 countries. ISO's central secretariat (located in Geneva, Switzerland) coordinates the system.

Micro Browsers: Micro browsers are browsers for mobile devices. These browsers are capable of interpreting WML instructions and executing WMLScript code.

Ubiquitous Computing: Refers to the ability to perform computing or communications from anywhere at any given time. Thus, untethering individuals from wired networks creates opportunities but is constrained by other issues such as power consumption.

WAP: Wireless Application Protocol (WAP) is the leading application standard to deliver information for wireless devices, such as cell phones. WAP is similar to the HTTP (Hypertext Transport Protocol) for the Web, and it is based on other Internet standards such as HTML, XML, etc.

WML: Wireless Markup Language (WML) is a markup language inherited from HTML and XML. WML is used to create Web pages specifically for micro browsers in mobile devices.

WMLScript: Scripting for micro browsers. WMLScript is used with WML to offer some dynamic effects on WAP Web pages.

User–Centered Mobile Computing

Dean Mohamedally
City University London, UK

Panayiotis Zaphiris
City University London, UK

Helen Petrie
City University London, UK

INTRODUCTION

Mobile computing and wireless communications continue to change the way in which we perceive our lifestyles and habits. Through an extensive literature review of state-of-the-art human-computer interaction issues in mobile computing (Mobile HCI), we examine recent pertinent case studies that attempt to provide practical mobile capabilities to users. We thus contribute to the reader a primer to the philosophy of developing mobile systems for user centred design.

User centred design elicits the needs and requirements of end users. Its purpose in mobile systems is to enable useful computing and communicating experiences for diverse types of users, anywhere at anytime and on demand. We shall therefore illustrate to the reader some of the key constraints of mobile devices such as limited visuals, contextual awareness and mobility itself, and more importantly how they can be overcome through innovative design and development.

INFORMATION VISUALISATION

One of the most fundamental objectives in the miniaturisation of computer technologies is to present a platform from which users can maintain usable levels of interaction with their data from wherever they are. Information visualisation has come a long way since the days of two-colour text-only format screens. Yet constraints determined by factors of physical engineering feasibility, such as screen quality and resolution, battery longevity and network capabilities, give us a particularly popular arena for exploration in mobile HCI research.

Such constraints are being addressed in a number of novel interfaces, such as Electronic Ink-based screens (E-Ink, 2004) that have a high resolution similar to that of natural paper, have low power requirements, and will eventually be capable of being rolled up for storage. TabletPCs (Microsoft, 2003) have the capability to use ultra-low powered pressure-sensitive pen input devices to elicit the amount of pressure incurred and also capture motion gestures on a visual interactive user interface, creating a sense of depth perception visualisation. They also enable very distinctive handwriting recognition without the need for learning preset handwriting letter shapes.

The Rapid Serial Visual Presentation (RSVP) concept (Bruijin, Spence, & Chong, 2001; Goldstein, Oqvist, Bayat-M, Ljungstrand, & Bjork, 2001) is one of many research investigations into methods of presentation of information on a small screen (Jones, Buchanan, & Thimbleby, 2002). By the beginning of 2001, over 88 million WAP (Wireless Access Protocol) hits on mobile phone screens were made in the UK alone (WAP Forum, 2003), and therefore significant effort has been undertaken on design factors of WAP site browsing globally.

LOCATION AND GEOGRAPHIC AWARENESS

Much of the research in geographic and location aware mobile systems correlate with mixing user requirements in information visualisation with geographic sensors. For example, audio user interfaces

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

for Global Positioning Satellites (GPS) receivers have typically been designed to meet the needs of visually handicapped users by giving audible signals as they travel past real-world coordinates that have been pre-designated in their system.

The proliferation of GPS-based location services has seen an impressive array of uses for location awareness in mobile deployment activities. They are now becoming an embedded part of upcoming generations of mobile phones and smart personal devices in mass consumer products. Examples include experimental tourist guides (Cheverst, Mitchell, & Davies, 2001; Davies, Mitchell, Cheverst, & Blair, 1998), navigation systems for disabled and elderly people (Holland & Morse, 2001; Petrie, Furner, & Strothotte, 1998), and also in location aware collaborative systems (Rist, 1999).

Close range location awareness technologies include Bluetooth (1998) based and RFID—Radio Frequency Identification (2003) based devices with which broadcast points can beam radio data to compatible handheld receivers. The results of current research in this domain is leading to opportunities in a variety of user scenarios that are made aware of your unique presence, for example, walking into a personally aware room will turn on the lights at your chosen settings, or location enabled advertising billboards will be able to read your public profile and communicate suitable electronic media to you wirelessly.

CONTEXTUAL AWARENESS

An issue in HCI research is investigating models and scenarios for maintaining consistency between the user's understanding of their environment, the understanding of the environment reported by the system and the actual state of the environment. Context-aware systems have to react not only to the user's input but also input (i.e., context) from the user's environment (Brown, Bovey, & Chen, 1997; Schilit, Adams, & Want, 1994). This offers opportunities for helping people to accomplish their goals effectively by understanding the value of information.

A realisation in this domain involves a specifically mobile systems orientated question—do we push the contextual information into the mobile system as they move within monitored zones or pull it on demand at their request? One of the challenges in context awareness is discovering computing services and resources available in the user's current environment, utilising discovery protocols such as Jini (2001) and SDP by Czerwinski, Zhao, Hodes, Joseph, and Katz (1999).

SENSORY-AIDED MOBILE COMPUTING

Mobile HCI has also changed the nature of computing for the demographics of users that have sensory disabilities, or alternatively require sensory enhancement. A low visibility prototype with supplemented tactile cues is presented in Sokoler, Nelson, and Pedersen (2002) with the TactGuide prototype. This is operated by subtle tactile inspection and designed to complement the use of our visual, auditory and kinesthetic senses in the process of way finding. It was found to successfully supplement existing way finding abilities. A mobile system that lets a blind person use a common laser pointer as a replacement of the cane in demonstrated by Fontana, Fuiello, Gobbi, Murino, Rocchesso, Sartor, and Panuccio (2002), who presented an electronic travel aid device that enables blind individuals to "see the world with their ears." A wearable prototype was constructed using low-cost hardware with the ability to detect the light spot produced by the laser pointer. It would then compute its angular position and depth, and generate a correspondent sound providing the auditory cues for the perception of the position and distance of the pointed surface. Another wearable system for blind users to aid in navigation is presented by Petrie et al. (1998), which projects a simple visual image in tactile form on the back or stomach.

Aside from aiding those disabilities it should be noted that sensory enhancement is an area for particular growth in Mobile HCI. Mobile systems that can augment the senses such as vision with heat and electrical sensors and sonic receivers are all ideas that can be investigated with undoubtedly a wide arena of scenarios.

COLLABORATIVE SYSTEMS

Mobile systems in general are becoming refined as instruments for co-operative wireless computing com-

munications in various forms of collaborative HCI. One successfully enhanced scenario is presented with air traffic controllers by Buisson and Jestin (2001). They constructed an effective solution for a distributed interaction prototype that would assist desk-based systems with the collaboration of a mobile operations manager in a fast paced and safety critical environment. The ability to work in mobile teams simulating CSCW (Computer Supported Collaborative Work) models is an important consideration for time critical and location dependent processes.

Collaboration is also an extensive area in augmented reality systems like that found in Nigay, Salambier, Marchand, Renevier, and Pasqualetti (2002). Augmented reality systems overlay information visualisation on top of physical views of the real world. Here it addresses the combination of the physical and digital worlds seamlessly in the context of a mobile collaborative activity. Mobile collaboration also takes place in VNC (Virtual Network Computing) based solutions for mobile devices such as PalmVNC for Palm PDAs (Minenko, 1998). VNC technology allows several users to view a desktop remotely and may allow them access to basic interactivity. This has a significant scope for future research as a remote-access collaboration method.

HOME AND DEVICE CONTROLLERS

Much of the research in the area of mobile human-computer interaction has focused on the user interfaces to the mobile devices themselves such as their input methods and displays. We can envisage how devices can fit into user-centred domains of information and control space.

Mobile device controllers for the home are still currently a relatively new and underdeveloped area for consumer interests beyond the usual remote controls (Weiser, 1991). We find now that protocols are being developed by companies and peer driven international committees, such as those that will enable the future wave of mobile and non-mobile devices to co-operate together on much more ubiquitous levels and facilitate growth in this area. Examples of these include Bluetooth short range wireless networking, Jini (2001) embedded devices networking in everyday home consumer equipment like light switches and kettles, and HAVi (2000) home media connectivity networking for audio visual equipment, to name a few.

An example of modelling future simulations of device controllers is found in Huttenrauch and Norman (2001), where a device is simulating the control of household robots. A popular consumer orientated protocol being corporately developed is the X10 (2000) protocol, which can control and relay electrical hardware information to other protocols such as e-mail and Internet connectivity. This combined with mobile scenarios can give rise to a host of ubiquitous and ambient intelligent hardware in physical locations. For example, one may wish to turn on their house lights, open the car garage and start the kettle boiling remotely with a single text message home to an X10 driven mobile interfacing system.

SOCIOLOGICAL VIEWS

Sociological aspects of mobile HCI are changing the way we live our everyday lives beyond our desktop computers at work and at home and there are new paradigms forming in these areas. For example, social communication by SMS text messaging using innovative short hand letter sequences rather than usual language syntax, and the development of assistive technologies such as Tegic Corp's patented T9 (1997) text prediction algorithm.

SMS text messaging and mobile chat messaging has changed our dimension of communications, despite its weak and sometimes unreliable connectivity. It is an asynchronous channel of communication that operates upon a principle of "store and forward": the sender sends a message when his or her device has a connection, and the message is forwarded to the recipient when the recipient's device has a connection. Given its asynchronous nature, we find that it is less obtrusive than real-time communications to utilise and respond to, as users may reply at their discretion. This contrasts with traditional voice telephony over mobiles—a synchronous channel of communication, which requires both mobile devices to establish a connection simultaneously.

The future of text messaging has been described as the advent of picture and video messaging and streaming, and it is already becoming apparent that its usage patterns are changing our cultural per-

spectives. In some countries it is already banned for religious, security, and privacy reasons as any unsuspecting person or entity can be the subject of a discrete imaging mobile device. Text messaging in combination with instant messaging and blogging (an Internet-viewable personal diary) techniques will present interesting dimensions to our social patterns.

MOBILE-BASED LEARNING SYSTEMS

M-learning systems that utilise mobile technologies and models of ubiquity are an area for growth in mobile HCI, though popular in their own right. Primarily considered to be a classical model of knowledge presentation in mobile and wireless classroom scenarios, the blackboard model has had numerous developments to enhance the capabilities of electronic learning, such as Chang and Sheu's (2002) ad hoc classroom system which enables students to migrate their daily activites to PDAs for digital recording of all of their events and contributions.

Learner-Centred Design (LCD) is an approach to building software that supports students as they engage in unfamiliar activities and focuses on enabling them to learn about a new area. LCD has been successfully used to support students using desktop computers for a variety of learning activities, and in Luchini, Quintana, and Soloway (2002), LCD is extended to the design of educational software for handheld computers. Here they presented a case study of ArtemisExpress, a tool that supports learners using handheld computers for online research. The results from this demonstrate that while user-centred design methods typically focus on software to support the work of expert computer users, LCD techniques in mobility focuses on directly providing learners with the educational support needed to learn about the content, tasks, and activities of the new domain they are exploring at their own pace and in their own environments.

NAVIGATION AND READABILITY

Voice recognition and synthesis has come an impressive way in Computer Science. Motorola's Mya Voice Browser as described by Chesta (2002) uses Automatic Speech Recognition (ASR) to understand and process human speech, capture requested information from voice-enabled Web sites, and then deliver the information via pre-recorded speech or Text-To-Speech (TTS) synthesis software that "reads" the relevant data to the user. The issues of internationalisation with such a system are covered in their research.

Researchers have to remain critical of the choices of navigation design, as found in Chesta (2002). An important requirement in HCI evaluations is to derive the usability and functional accuracy of a designed system in its domain. As Chittaro and Cin (2001) found in their results, the WAP/WML protocol navigation capabilities needs to be reviewed for user performance, in particular they observed the WAP methods for navigating links, list of links, action screens and selection screens.

GRAPHICS ENGINEERING

Computer graphics engineering associates closely with HCI, especially where constraint user interfaces are concerned. In general, researchers have been trying to find the most perceptually accurate and aesthetically pleasing representations for allowing humans to access and visualise computational information as responsively as possible. Several distinct disadvantages of current day mobile systems is demonstrated by the lack of screen space available and the hardware demands of the limited processing and power usage capabilities available to generate fast computer graphics. As technologies and industry standards develop however, some of these constraints will be removed altogether.

Constraint visual computing experiences are pushing mobile user interface requirements into constructing new and more powerful miniature hardware and software for the support of video and real-time 2D/3D acceleration. This can be seen in the impressive work by the Khronos group (2002) to construct an open platform for mobile graphics software developers. NVidia Corp's mobile embedded graphics processors (2004) augment this with new research opportunities by creating platforms for embedded real-time 3D computer graphics processing on mobile devices.

CONCLUSION

We have presented a review of current state of the art issues in user centred mobile systems that researchers have been involved in, and explored some of their solutions. In this youthful field, research that has published applications and techniques from the period 1997 – 2004 are going on to influence research directions in the field of mobile HCI and have been described in this section as several key categories of computer science.

The future of mobile HCI research holds great promise in the culmination of the user centred design issues as noted, which will lead into the field of ubiquitous computing. This will stem from trends in digital capture, processing and presentation of real-time and real-world data that is embedded in our environment. For users of mobile systems, it gives hope to the development of systems that will one day provide tools that can react, adapt and assist our dynamic lifestyles and enhance both our naturally individual and collaborative ways of life.

REFERENCES

Bluetooth official membership site (1998). Retrieved on March 14, 2004, from *http://www.bluetooth.org*

Brown, P.J., Bovey, J.D., & Chen, X. (1997). Context-aware applications: From the laboratory to the market place. *IEEE Personal Communications, 4*(5), 58-64.

Bruijin, O.D., Spence, R., & Chong, M.Y. (2001). RSVP browser: Web browsing on small screen devices. *ACM Mobile HCI Workshop 2001*, Lille, France.

Buisson, M. & Jestin, Y. (2001). Design issues in distributed interaction supporting tools: Mobile devices in an ATC working position. *ACM Mobile HCI Workshop 2001,* Lille, France.

Chang, C.Y. & Sheu, J.P. (2002). Design and implementation of Ad Hoc classroom and eSchoolbag systems for ubiquitous learning. *IEEE WMTE 2002*.

Chesta, C. (2002). Globalization of voice-enabled Internet access solutions. *ACM Mobile HCI Conference 2002*, Pisa, Italy.

Cheverst, K., Mitchell, K., & Davies, N. (2001). Investigating context-aware information push vs. information pull to tourists. *ACM Mobile HCI Workshop 2001*, Lille, France.

Chittaro, L. & Cin, P.D. (2001). Evaluating interface design choices on WAP phones: Single choice list selection and navigation among cards. *ACM Mobile HCI Workshop 2001*, Lille, France.

Czerwinski, S., Zhao, B., Hodes, T., Joseph, A., & Katz, R. (1999). An architecture for a secure service discovery service. *Proceedings of MobiCom '99*, Seattle, Washington, August.

Davies, N, Mitchell, K., Cheverst, K., & Blair, G. (1998). Developing a context sensitive tourist guide. *ACM Mobile HCI Workshop 1998*, Glasgow, UK.

E-Ink (2004). Retrieved on March 14, 2004, from *http://www.eink.com*

Fontana, F., Fuiello, A., Gobbi, M., Murino, V, Rocchesso, D, Sartor, L., & Panuccio, A. (2002). A cross-modal electronic travel aid device. *ACM Mobile HCI Conference 2002*, Pisa, Italy.

Gershon, N., Card, S., & Eich, S.G. (1997). Information visualization. *Chi '97 Tutorial Notes*, Atlanta, Georgia, March 22-27.

Goldstein, M., Oqvist, G., Bayat-M, M., Ljungstrand, P., & Bjork, S.(2001). Enhancing the reading experience: Using adaptive and sonified rsvp for reading on small displays. *ACM Mobile HCI Workshop 2001*, Lille, France.

HAVi Consortium. Retrieved on March 14, 2004, from *http://www.havi.org*

Holland, S. & Morse, D.R. (2001). Audio GPS: Spatial audio in a minimal attention interface. *ACM Mobile HCI Workshop 2001*, Lille, France.

Huttenrauch, H. & Norman, M. (2001). PocketCERO - Mobile interfaces for service robots. *ACM Mobile HCI Workshop 2001*, Lille, France.

Jini - The community resource for Jini technology (2001). Retrieved on March 14, 2004, from *http://www.jini.org/*

Jones, M., Buchanan, G., & Thimbleby, H. (2002). Sorting out searching on small screen devices. *ACM Mobile HCI Conference 2002*, Pisa, Italy.

Khronos Working Group. (2004). Retrieved on March 14, 2004, from *http://www.khronos.org*

Luchini, K., Quintana, C., & Soloway, E. (2002). ArtemisExpress: A case study in designing handheld interfaces for an online digital library. *ACM Mobile HCI Conference 2002*, Pisa, Italy.

Minenko, V. (1998). PalmVNC. Retrieved on March 14, 2004, from *http://www.wind-networks.de/PalmVNC/*

Nigay, L., Salambier, P., Marchand, T., Renevier, P., & Pasqualetti, L. (2002). Mobile and collaborative augmented reality: A scenario based design approach. *ACM Mobile HCI.*

NVidia Corp. (2004). Retrieved on March 14, 2004, from *http://www.nvidia.com*

Petrie, H., Furner, S., & Strothotte, T. (1998). Design lifecycles and wearable computers for users with disabilities, *ACM Mobile HCI Workshop 1998*, Glasgow, UK.

RFID Organisation (2003). Retrieved on March 14, 2004, from *http://www.rfid.org*

Rist, T. (1999). Using mobile communication devices to access virtual meeting places. *ACM Mobile HCI Workshop 1999*, Edinburgh, UK.

Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. *Proceedings of the Workshop on Mobile Computing Systems and Applications*, Santa Cruz, California.

Sokoler, T., Nelson, L., & Pedersen, E.R. (2002). Low-resolution supplementary tactile cues for navigational assistance. *ACM Mobile HCI Conference 2002*, Pisa, Italy.

TabletPC (2003). Microsoft Corp. Retrieved on March 14, 2004, from *http://www.microsoft.com/tabletpc*

Tegic Corp T9 text prediction system. (2004). Retrieved on March 14, 2004, from *http://www.tegic.com*

WAP Forum press release (2004). Retrieved on March 14, 2004, from *http://www.wapforum.org/*

Weiser, M. (1991). The computer for the 21st century. *Scientific American,* September, 09-91.

Weiss, S. (2002). *Handheld usability.* UK: Wiley Publishing.

X10 electrical-to-computer-software hardware protocol and controller equipment (2004). Retrieved on March 14, 2004, from *www.x10.com*

KEY TERMS

CSCW: Computer-supported collaborative work, a discipline of computer science dedicated to the use of computer tools to allow groups of participants to work together in the resolution of a problem domain.

Information Visualisation: A process of transforming information into a visual form enabling the user to observe information (Gershon, Card, & Eich, 1997).

M-Learning: The use of mobile devices as tools in the computer science discipline of electronic learning.

Mobile HCI: The Human Computer Interaction (HCI) aspects of the design, evaluation and application of techniques and approaches for all mobile computing devices and services. (ACM Mobile HCI, 2001).

PDA: Personal Digital Assistant; a handheld computing device that may contain network facilities but generally is used for personalised software purposes beyond a standard organiser.

RFID: Radio Frequency Identification; a technology that uses radio frequency waves to communicate data between a moveable item with a small digital tag and a reader to identify, track, or locate that item.

Ubiquitous Computing: The evolution of mobile HCI whereby user centred principles of hardware and software development embed the nature of mobile computing into the background of everyday life.

WAP: Wireless Access Protocol; a protocol for implementing advanced telecommunications services for accessing Internet pages from mobile devices.

WML: Wireless Markup Language; based on the XML language it has been derived to create a user interface and content specification for WAP-enabled devices. (WAP Forum 2003).

Using Semantics to Manage 3D Scenes in Web Platforms

Christophe Cruz
University of Bourgogne, France

Christophe Nicolle
University of Bourgogne, France

Marc Neveu
University of Burgundy, France

INTERNET AND 3D SCENES

Computer graphics has widely spread out into various computer applications. After the early wire-frame computer generated images of the 60s, spatial representation of objects improved in the 70s with Boundary Representation (B-Rep) modeling, Constructive Solid Geometry (CSG) objects, and free-form surfaces. Realistic rendering in the 90s, taking into account sophisticated dynamic interactions (between objects or between objects and human actors, physical interactions with light, etc.) now make 3D-scenes much better than simple 3D representations of the real world. Indeed, they are a way to conceive products (industrial products, art products, etc.) and to modify them over time, either interactively or by simulation of physical phenomena (Faux and Ellis Horwood Ltd. & Pratt, 1979; Feiner & Foley, 1990; Mortenson, 1985).

The exponential development of Internet tends toward two domains which may seem contradictory. On the one hand, we note the increasing importance of the visual aspect of the Web inasmuch as the text that initially composed pages of the first Web sites has been replaced by pictures and animation. We note the breakthrough of software such as Flash from Macromedia (2004) in this domain. On the other hand, the informative aspect of the Web has undergone major development with the interconnection of databases with HTML pages using ASP, PHP, and so on. Such information becomes intelligent, adaptived to the behavior of the connected users. The breakthrough in the interconnection of databases with the HTML pages has permitted the creation of new dynamic sites. Today, a Web site must be lively, attractive, intelligent, active, and interactive.

Nevertheless, limits do exist. In terms of the visual aspect, 3D representation on Internet is expanding rapidly. However, it is often limited to short animated sequences short animated sequences, due to the important resources needed to use 3D on the network. In terms of the informative aspect, it is often limited to the interfacing of database with the HTML code.

Large amounts of data can be generated from such variety of 3D-models. Because there is a wide renge of models corresponding to various areas of applications (metallurgy, chemistry, seismology, architecture, arts and media, etc.) (The DIS 3D Databases, 2004; The Fermi Surface Database, 2004), data representations vary greatly. Archiving these large amounts of information most often remains a simple storage of representations of 3D-scenes (3D images). To our knowledge, there is no efficient way to manipulate, that is archive, extract, and modify, scenes together with their components. These components may include geometric objects or primitives that compose scenes (3D-geometry and material aspect), geometrics transformations to compose primitives objects, or observation conditions (cameras, lights, etc.). Difficulties arise less in creating 3D-scenes, rather than in the interactive reuse of these scenes, particularly by database queries, for example, via the Internet. Managing 3D-scenes (e.g., querying a database of architectural scenes by the content, modifying given parameters on a large scale, or performing statistics) remains difficult. This implies that DBMS should use the data structures of the 3D-scene models.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Unfortunately, such data structures are often of different or exclusive standards. Indeed, many standards exist in computer graphics. They are often denoted by extensions of data files. Let us mention, as examples, 3dmf (Apple's Quickdraw 3D), 3ds (Autodesk's 3D-Studio), dxf (AutoDesk's AutoCAD), flt (Multigen's ModelGen), iv (Silicon Graphics' Inventor), obj (Wavefront/Alias), and so on. Many standardization attempts strive to reduce this multiplicity of various formats. In particular, there is STEP (Standard for the Exchange of Product model data), an international standard for computer representation and exchange of products data (Fowler, n.d.). Its goal is to describe data bound to a product as long as it evolves, independently of any particular computer system. It allows file exchanges, but also provides a basis for implementing and sharing product databases. Merging 3D information and textual information allows the definition of the project's mock-up. Indeed, 3D information describes CAD objects of the projcet and textual added information gives semantic information on geometries. The main issues are the sharing and the exchange of the digital mock-up. The next section explains how we use a digital mock-up to create an information system with the help of the semantic included in geometric information. Information is exchanged and shared through a Web platform.

BACKGROUND

With the emergence of new powerful computers, the 3D models created by computer-aided design tools are huge and very complex. The plans of a boat, plane, or architectural structure can exceed a gigabyte in size. The GigaWalk (Baxter, Sud, Govindaraju & Manocha, 2002) project is a rendering system making it possible to display projects of CAD with more than 10 million polygons. The design based on the simulation of these data cannot make a useful contribution without the possibility of generating an interactive display through a virtual visit of the model. Many optimizations and acceleration techniques for interactive display were developed for this type of data. These techniques include visibility computations, object simplification and image-based representation. All these techniques have been combined successfully in the rendering of specific data including architectural models (Funkhouser, Teller, Sequin & Khorramabadi, 1996) and urban models (Wonka, Wimmer & Sillion, 2001). The digital mock-up greatly impacts the financial and strategic choices of companies during the design phase. To improve the quality of prototyping and refined strategic choices, collaborative platforms were developed on the Web. Along with digital mock-up, these platforms allow designers and decision maker architects to work directly with geographically distant companies (Torguet, Balet, Gobbetti, Jessel, Duchon & Bouvier, 1999).

Nevertheless, these collaborative platforms do not allow the geometrical handling of a great quantity of polygons in real time without a prohibitory precalculates time. A way to solve this problem is to structure the 3D scene according to semantic criteria or to start from the only geometrical criteria only. Semantics is a crucial point for Web platforms because it influences the three characterizing axes of platforms, namely data, communication, and processes.

- Data is the information which is handled through the system. This information includes the data from the digital mock-up, the data of concerning model management like users and rights associated with users, and a set of meta-data allowing data management on a higher level of abstraction. This level allows the handling of the semantics of information and thus making the information more relevant to the situation of the user.
- Communication is the infrastructure which is installed to transfer information between processes and project actors. Transfer of more relevant information will limit the size of information exchanged and thus will improve the response times in the communications between processes.
- Processes carry out actions which are ordered either by another process or by an actor of the project. Processes are either generic or specialized. A set of generic processes forms the core of the system, making it possible to carry out simple actions which correspond to the use context of the platform. Specialized processes are composed of a sequence of simple processes and specialized processes to undertake a complex action. For example, a simple process will make it possible to insert an individual into a database and a complex process will make it

possible to insert a hierarchy of individuals into a database. This specialized process uses two simple processes: the insertion of a person and the creation of a hierarchy link between two people in the database.

The next section describe a emerging approach, ACTIVe3D, and the influence of semantic on the three characterizing axes of a Web platform.

A NEW APPROACH

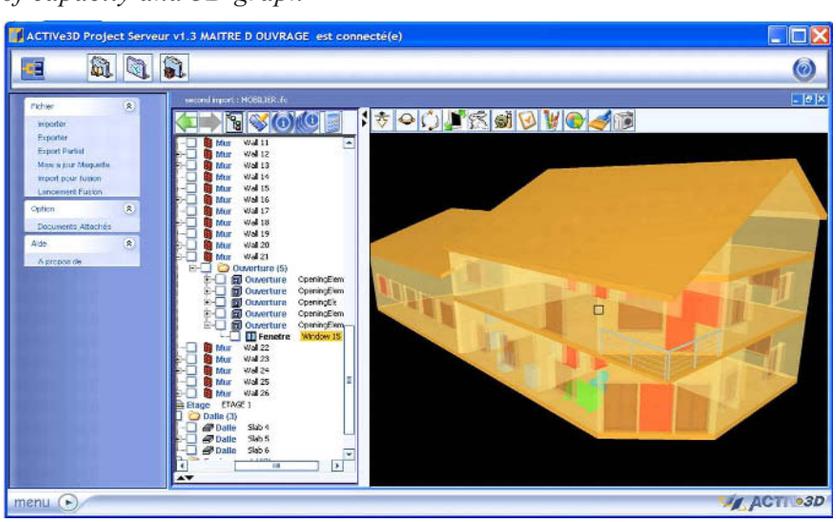
The ACTIVe3D method proposes a solution which makes it possible to associate semantics trade with the objects represented in complex geometrical 3D models. This association provides contextual trees which associate dynamically, using rules, a knowledge trade with groups of polygons to generate 3D trade objects. The dynamic feature of this method relates to the automatic generation of the 3D scene starting from the CAD files and the possibility of handling each trade object in the scene and of associating specific documents or functions with it. For example, it is very useful to select a door in a 3D scene to obtain the corresponding invoice and/or to activate the Web service which provides a description of the product based on the catalogue of the supplier. The use of contextual trees limits the geometrical complexities of the model which enables its use on Internet within reasonable deadlines.

To do this, ACTIVe3D is based on an ontology, which specifies semantic information contained in project information. This ontology defines vocabulary, concept, and relationships for the manipulation of hierarchical data. It also defines a formal framework to manage information according to its semantics, not only according to structure, thus allowing the dynamic creation of 3D scenes by context. The context view of the 3D scene is a trade association view, which shows a relevant view of the digital mock-up. Thanks to this new abstract level, the three axes of Web platforms are redefined.

The data axis uses industrial foundation classes (IFC), which is an ISO norm that defines all components of a building in a civil engineering project. IFC files are textual files whose size can reach 100 megabytes. Several IFC files can coexist on the same civil engineering project. Due to their size, their handling and sharing is a complex task. An IFC file for a standard building can contain more than 300 000 business objects organized in a cyclic graph. Each node of the graph includes partial semantic information. To obtain complete semantic information on a trade object, it is necessary to analyze several nodes which are not inevitably directly dependent. To address this problem we have developed a methodology based on graph analysis and tree classification. This methodology is articulated in two steps.

The first step is an analysis and conversion of each object and connection from the source file into

Figure 1. IFC tree of capacity and 3D graph



acyclic graphs called contextual trees. This process is undertaken using business rules. An example of a business rule is: a window is in an opening element in a wall. The main tree resulting from this process, is the geometrical contextual tree which contains the topological relations between the various objects. Other contextual trees are built starting from the IFC files, such as the contextual tree of capacity defining object composition (a building contains two floors, a floors contains beams, walls, and so on.) This step is completed when all information contained in the source IFC files are represented in contextual trees. Figure 1 displays a snapshot containing the view of a capacity tree and geometrical tree.

The second step is dedicated to 3D modeling (Abrams, Watsen & Zyda, 1998; Bowman, Davis, Badre & Hodges, 1999; Kim, Hwang & Kim, 2002; Szabo, Stucki, Aschwanden, Ohler, Pajarola & Widmayer, 1995). In the 3D scene generation process, all the geometry contained in IFC files are converted into triangular surface model (Ronfard & Rossignac, 1996). During this conversion, the 3D objects are associates with the GID. The GID is the general identifier used to identify each business object in an IFC file. This GID is used to link the 3D visualization with the information stored in the databases. All insertion of new data in any base is referenced by a GID corresponding to an IFC object. All trees generated in the platform are XML trees. We have developed a specific database schema dealing with the semantic and the 3D aspects of the IFC. This schema is based on the ACTIVe3D ontology. The trees and the component elements are stored in a relational database and manipulated using SQL. From this database and the GID, all types of information can be attached to the 3D visualization of a business object.

The communication axis will be adapted to facilitate the exchange of information through the network. All trees generated in the platform are XML trees, so Web services give us a framework to easily carry XML information with the help of HTTP network level. The data flows use Web services, but we have also defined an internal structure of information exchange. Indeed, the ACTIVe3D architecture based on a central router, allows each specialized module to exchange and co-operate in order to answer user queries. The database module contains a set of processes that allows the inter-operability of several local and distant databases. The GED modules allow the user to associate documents attached to the 3D objects. The other modules developed in the ACTIVe3D platform concern specific business processes from civil engineering. The Web services contained in each module can be combined by the router to resolve user queries.

The processes axis uses the contextual trees. Each functional unit and each context are manipulated as XML documents through web services. The document can be converted into IFC in the output of the system. This conversion allows the civil engineering participants to exchange maps throughout the life cycle of a civil engineering project. IFC Services provided on the ACTIVe3D server are XSL processes associated with a context. The use of XSL is extended to generate other documents such as technical reports and so on. In the same way, the graphic contextual trees are transformed into X3D documents. X3D is an XML language for the description of 3D scenes. Thus, the 3D scene is customized according to the

Figure 2. A 3D scene in a plumbing context

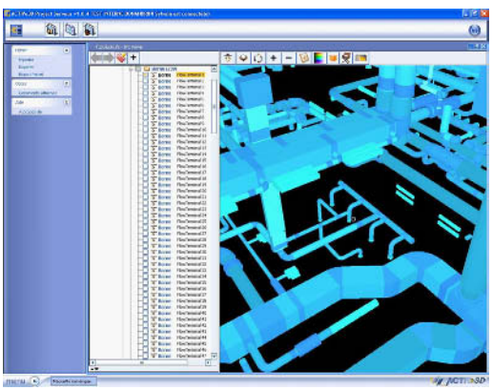
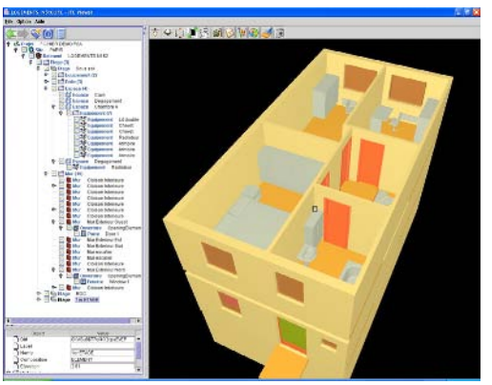


Figure 3. A 3D scene in an architectural context



service concerned. For example, as Figures 2 and 3 show, it is possible to generate two different 3D scenes from two differents contexte trees in real time resulting from two the same information source. Moreover, the graphic elements preserve connections with corresponding trade objects stored in the databases. Thus, the data contained in the systemes can be manipulated from the 3D scenes or from the context trees.

CONCLUSION

This article presents the technical evolution in 3D manipulation and storage. Currently, research in this domain concerns the combination of semantics with 3D representation. The focus of this article concerns the fact that the semantic approach is useful in managing 3D scenes in business environments. Indeed, semantics allows users to extract relevant information from a relational database depending on the context. Moreover, extracted information is less complex then the complete model. Semantics therefor helps us validate partial extracted information from our information system. Thus, semantics expressed by textual information validates geometrical information. For example: We cannnot put a wall in a door, and the capacity graph validates this information into memory or from the database by means of simple process. The semantic approach also allows users to validate more complex information that combines textual and geometrical information. Indeed, semantics is a powerful tool in computer aided mistake detection in engineering projects and in 3D scenes as well. Imagine that we want to discover if a wood beam "kind x232" cross a load-bearing wall. First, we have to extract the scene graph of bounding box with only business objects as wood beam and load-bearing wall. Next, we start a process to detect if a bounding box of a wall in the scene graph crosses a bounding box of a load-bearing wall. In this example, we note the two steps required to arrive at detection. The first one consists in defining which kind of information we need to create a dynamical graph. This graph will contain all information required in the second step. The second step consists in defining rules concerning business objects and relation elements with the help of mathematical logic. The result of the mistake detection will use data from both steps.

Our future work will consist of the creation of a formal framework allowing users to define customized rules for mistake detections. This framework will help users to define each kind of business object managed and each logical rule need to extract mistakes.

REFERENCES

Abrams, H., Watsen, K., & Zyda, M. (1998). Three-tiered interest management for large-scale virtual environments. *VRST,* 125-129.

Baxter III, W.V., Sud, A., Govindaraju, N.K., & Manocha, D. (2002). GigaWalk: Interactive walkthrough of complex environments. *Eurographics Workshop on Rendering.*

Bowman, D., Davis, E., Badre, A., & Hodges, L. (1999). Maintaining Spatial Orientation during Travel in an Immersive Virtual Environment. *Teleoperators and Virtual Environments, 8*(6), 618-631.

The DIS 3D databases (2004). Online *http://dtp.nci.nih.gov/docs/3d_database/dis3d.html*

Extensible Markup Language (2004). Online *http://www.w3.org/XML/*

Faux and Ellis Horwood Ltd. & Pratt, J. (1979). *Computational geometry for design and manufactures*.

The Fermi Surface Database (2004). Online *http://www.phys.ufl.edu/fermisurface/*

Foley, J., vanDam, A., Feiner, S. & Hughes, J. (1990). *Computer graphics: Principles and practice*. Addison-Wesley.

Fowler, J. (n.d.). *Step for dated management exchange sharing and technology appraisals*. Twickenham, Middlesex, UK.

Funkhouser, T., Teller, S., Sequin, C., & Khorramabadi, D. (1996). The UC Berkeley system for interactive visualization of large architectural models. *The Journal of Virtual Reality and Teleoperators, 5*(1), 13-44.

IAI (2004). Online *http://www.iai-international.org/iai_international/*

Kim, B.H., Hwang, J., & Kim, Y.C. (2002). The design of high-level database access method in a Web-based 3D object authoring tool. *The Fourth International Conference on Distributed Communities on the Web*, April 3-5, Sydney, Australia.

Macromedia (2004). Online *http://www.macromedia.com/*

Mortenson, M.E. (1985). *Geometric modeling*. Wiley.

Ronfard, R. & Rossignac, J. (1996). Full-range approximation of triangulated polyhedra. *Computer Graphics Forum, 15*(3), 67-76.

Szabo, K., Stucki, P., Aschwanden, P., Ohler, T., Pajarola, R., & Widmayer, P. (1995). A virtual reality based system environment for intuitive walk-throughs and exploration of large-scale tourist information. *Proceedings of the Enter95 Conference* (pp. 10-15).

Torguet, P., Balet, O., Gobbetti, E., Jessel, J.P., Duchon, J., & Bouvier, E. (1999). Cavalcade: A system for collaborative prototyping. In G. Subsol (Ed.), *Proceedings of the International Scientific Workshop on Virtual Reality and Prototyping*, Laval, France, June 3-4, (pp. 161-170).

Wonka, P., Wimmer, M., & Sillion, F. (2001.) Instant visibility. *EG 2001 Proceedings, 20*(3), 411-421.

KEY TERMS

B-Rep: In boundary representation (B-Rep), complex geometrical forms are described using their boundary surfaces. In this process, the surface of an object is broken down into smaller polygons, mainly triangles. This therefore makes this type of modeling particularly suitable for irregularly shaped surfaces. Most animation programs use this method.

CAD (Computer Aided Design): The use of computer programs and systems to design detailed two- or three-dimensional models of physical objects, such as mechanical parts, buildings, and molecules.

CSG: There are few ways to describe a three-dimensional model. One of the most popular is Constructive Solid Geometry (CSG). In CSG, a model is compiled from primitives and Boolean operators linking them. Data are stored in the tree structure, where the leaves are the primitives, and the nodes are the operations: intersection (AND), union (OR), and complement (NOT).

Cyclic Graph: A graph of n nodes and n edges such that node i is connected to the two adjacent nodes i+1 and i-1 (mod n), where the nodes are numbered 0, 1, ..., n-1. *http://mathworld.wolfram.com/CyclicGraph.html*

ISO Norm: "International Organization for Standardization" is a network of the national standards institutes of 148 countries, on the basis of one member per country, with a central secretariat in Geneva, Switzerland, that coordinates the system. ISO is a non-governmental organization. *http://www.iso.org*

XSL Style Sheet: XSL is a language for expressing style sheets. An XSL style sheet is a file that describes how to display an XML document of a given type. *http://www.w3.org/Style/XSL/*

Virtual Communities

George Kontolemakis
National and Kapodistrian University of Athens, Greece

Panagiotis Kanellis
National and Kapodistrian University of Athens, Greece

Drakoulis Martakos
National and Kapodistrian University of Athens, Greece

INTRODUCTION: THE EVOLUTION OF VIRTUAL COMMUNITIES

In recent years, computer mediated communication has been the enabling factor for connecting people to one another and establishing "virtual relationships" (Igbaria, 1999; Johnston, Raizada, & Cronin, 1996). Virtual communities evolved as users of the early networks utilized them mainly for informal rather than business-related communication. These communities were not planned development in the sphere of computer networking. As this form of interaction increased, the users began to demand better and improved technology and functionality which would assist them in their interactions. "Virtual Communities describe the union between individuals or organizations who share common values and interests using electronic media to communicate within a shared semantic space on a regular basis" (Schubert, 1999).

Four major milestones have marked the development and evolution of Virtual Communities. These are:

a. 1977 – Development of ARPAnet
b. 1978 – First Virtual Community (SF-LOVERS)
c. 1980 – USENET
d. 1990s – America Online (AOL)

The first virtual community was formed on ARPAnet as communication became easier due to the development and offering of more sophisticated functions (Cronin, 1995). Joseph C.R. Licklider and Robert Taylor, research directors for the US Department of Defense, started the research which lead to the development of ARPAnet in 1977; the first multisite, packet switched network. ARPAnet was designed to support the Advanced Research Projects Agency (ARPA) for the transferring of files and resource sharing. It was a simple services network for sharing news and for many to many synchronous communications. The two main features were the File Transfer Protocol (FTP) and TELNET, a remote login facility. E-mail was an afterthought in the development of ARPAnet, but quickly became one of the most popular features of the system. Once those were sufficiently developed, the necessary infrastructure and functionality was in place to enable the formation of a community. The first virtual community was Science Fiction Lovers (SF-LOVERS), started in 1978 (Cronin, 1995).

Many virtual communities followed. Starting in the early 1980s, a network called USENET was set up to link university computing centers that used the UNIX operating system. USENET came into being in late 1979, shortly after the release of V7 Unix with UUCP. Two Duke University graduate students in North Carolina, Tom Truscott and Jim Ellis, thought of hooking computers together to exchange information within the UNIX community. Steve Bellovin, a graduate student at the University of North Carolina, put together the first version of the news software using shell scripts and installed it on the first two sites: "unc" and "duke". At the beginning of 1980 the network consisted of those two sites and "phs" (another machine at Duke), and was presented at the January USENIX conference of the same year. Steve Bellovin later rewrote the scripts into "C" programs but those were never released beyond "unc" and "duke". Shortly thereafter, Steve Daniel did another implementation in C for public distribution. Tom Truscott made further modifications, and this became the "A" news release.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

One function of USENET was to distribute "news" on various topics throughout the network. Participants were able to set up their own "newsgroups" on topics of shared interest. These were bulletin board type discussions where participants could send messages to a newsgroup on a given topic and read the messages sent by others. Initially, all of the newsgroups focused on technical or scholarly subjects. Groups that focused on non-technical topics such as food, drugs, and music also started to appear. Before long, the number of newsgroups started to grow exponentially. From 158 newsgroups in 1984, the number grew to 1,732 groups in 1991 and to 10,696 groups in 1994. Today there are more than 25,000 different newsgroups in existence (Digital Places, 2003).

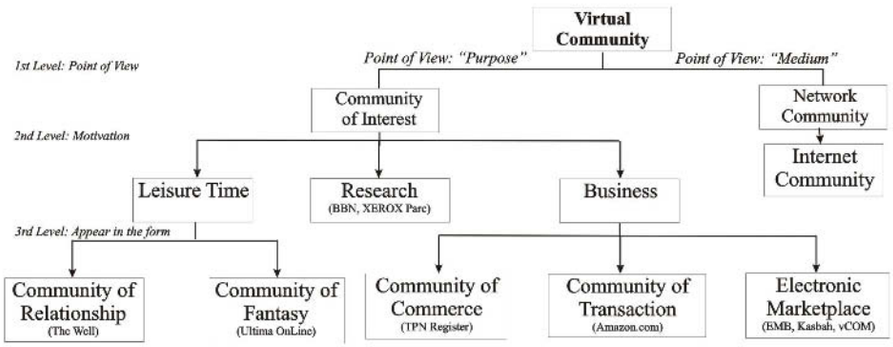
Commercial organizations began to take note of and exploit the trend. CompuServe hosted a number of "forums" that allowed people to share professional and personal interests and in 1980 was the first for providing real-time chat online as a service to its members. The popularity of these forums played an important role in the growth of CompuServe throughout the 1980s. In the early 1990s, AOL was establishing itself as an easy-to-use service for a mass audience. While it provided news and reference information and other kinds of services, AOL emphasized the value of person-to-person communication and the benefits of participating in virtual communities. AOL was in fact a portal to many popular online communities. Through AOL's site one could always find an online community that matched his personal interests. AOL provided communities for investors, cultures, pre-teenagers, and older adults. This was one of the factors that helped AOL became one of the largest Internet Service Providers (ISP).

CHARACTERISTICS AND TYPES OF VIRTUAL COMMUNITIES

According to Roberts (1998), there are six dimensions that characterize a community. The first one is *cohesion,* which is the sense of there being a group identity and that an individual belongs to the group. To achieve that, virtual communities must maintain the commitment of members for continuous participation and contribution through rituals and other practices that increase the individual's identification within the group. Small groups possess a special quality that enables them to maintain themselves with greater ease than larger groups. In particular, small groups are usually able to provide high levels of communication between each member of the group. The second one is *effectiveness,* which talks about the impact that the group has on the members' lives and the outside world. The community may be the primary vehicle for evolution in certain fields such as academia because various ideas and thoughts from any part of the world can help an issue or a program to evolve rapidly. The third one is *help,* which is the perceived ability of members to ask for and receive various types of assistance. The fourth one, *relationship,* is the likelihood of group members interacting individually, including forming friendships. This entails the emotional and affective bonds created between co-participants in a community. Group members can gradually form friendships when the community provides them with the means to share information, give financial support, attend conferences together, and so on. The fifth one is *language* and specifically the prevalence of a specialized language. Internet jargon and specialized language within the newsgroups are common. They are more likely on high-traffic lists, and, interestingly, on lists with large female membership. Finally, *self-regulation* refers to the ability of the group to police itself. This can be done by restraining and punishing individual actions that exploit or undermine collective goods through monitoring and sanctioning. Small groups maintain high levels of surveillance of each member's activities, especially his or her contributions and withdrawals to and from the group's resources.

A group of researchers at the Annenberg School of Communication at the University of Southern California identified four major components that contribute to creating a "sense of community": (a) need fulfilment, which shows how well a participant's needs are satisfied by a community; (b) inclusion, which shows the extent to which participants are open and encouraged to participate in each other's plans and activities; (c) mutual influence, which shows the extent to which participants openly discuss issues and affect one another; and (d) shared emotional experiences, which include sharing events that specifically arouse feeling and are typically memorable such as trips, birthdays, anniversaries, weddings, and so on.

Figure 1. A layered framework for categorizing virtual communities (adapted from Schubert and Ginsburg, 2000)



A categorization of virtual communities according to two different points of view, namely the underlying medium or the purpose they serve is depicted in Figure 1.

When the point of view is the medium, a community that is evolved through the classic sense of communication channels is the network community. A special case of a network community is an Internet community which evolves on the Internet and is not to be confused with the Internet community as a whole which is the total number of all Internet users. As stated in the definition given in the preceding section, communities are motivated by a common interest. In this context, on the first level, we also speak of "communities of interest". Communities of interest bring together participants who interact extensively about specific topics of interest (Armstrong & Hagel, 1997). Hence, depending on the interests of their members (social, academic, business, etc.), communities are distinguished as "leisure time communities", "research communities", and "business communities".

Leisure time communities are communities of people that are using computers for relaxation, fun and social interaction. They appear, at the third level, in the form of communities of relationship and communities of fantasy. Communities of relationship center on intense personal experiences and generally adhere to masking identities and anonymity (Armstrong & Hagel, 1997). Here, participants discuss the personal issues associated with these experiences and exchange information about support institutions. The Well (http://www.well.com), is a pioneering online community of relationship known for engaging conversation and intelligent debate. It regularly features more than 260 conferences ranging from technical and specific to abstract and surreal. There are also many communities of relationship that are designed to help people meet each other. For example, the "acmelove" Web site (http://acmelove.com) and the Michael Jackson Café (http://www.mjcafe.net/chat.htm) are two of these places (Koth, 2003). Communities of fantasy allow participants to create new personalities, environments, or stories of fantasy. Here, individuals can take on the persona of an imaginative or factual being and act out roles like members of a spontaneous improvisational theatre. Online gaming for instance is something that many users do in their leisure time. The Multi-Player Online Gaming Directory (http://www.mpogd.com/) for instance is an example of a site dedicated to providing information about multiplayer online games.

Research communities are characterized by formal frameworks for knowledge dissemination through communication and the sharing of opinions between their members (Crane, 1972). The following two are examples of research communities. The Concord Consortium (http://www.concord.org) is a non-profit educational research and development organization launched in 1994 by educators in Concord, Massachusetts and the XEROX Palo Alto Research Centre (PARC) (http://www.parc.xerox.com) carries out pioneering research that covers a broad spectrum of research fields ranging from electronic materials and devices through computer-based systems and software, to research into work practices and technologies in use.

Business Communities emerge within e-commerce environments. Their members have a com-

mercial interest and seek the relationship with business partners. Referring to Figure 1, communities of commerce, transaction, and marketplaces are forms of business communities. Communities of commerce describe business-to-business alliances between partners at all levels of the value chain with joint value creation as their aim. Two examples are CommunitiesofCommerce (http://www.communitiesofcommerce.com/) and the Sausalito.net commercial community (http://www.sausalito.net/commerce/). Communities of transaction deal with the exchange of goods and services, or more specifically the purchase transaction itself (Armstrong & Hagel, 1997). They can emerge between business-to-business partners as well as between companies and consumers. They provide a trustworthy commercial and social environment, mutual support and the means for the identification of individual user needs based on shared community knowledge.

Amazon's Recommendation Center (http://www.amazon.com), is a good example of a community of transaction. Amazon profits from consumers' and authors' input of reviews, recommendations, and additional information. This particular community has become quite large, giving economies of scope to the user community. Finally, electronic marketplaces are virtual communities where buyers and sellers exchange information, negotiate, and transact. Electronic marketplaces take various forms such as auctions, product exchanges, online shopping markets, e-catalogs, and so on. They represent one of the best examples for the evolution of the Internet; from a mere technical infrastructure to business enabler. A working example of this type of community is Kasbah (www.kasbah.com).

THE FUTURE: WEARABLE COMMUNITIES & VIRTUAL WORLDS

Mobile devices such as wearable computers and PDAs can help us form and maintain cooperative and interdependent relationships with the people we meet. Being small and unobtrusive, they will become our constant companions, ready and available wherever we go, and will be an accessory as common as glasses, a wallet, or keys. That means the learner can take opportunities directly in the situation where they occur, because he has his learning environment always at hand. With mobile devices, there would be a chance to put virtual post-its on any object, read post-its from others, and in the process become part of a location-aware community (Frohberg, 2003).

As we use mobile devices to store personal information about ourselves and others, our mobile companions become something more than just communication devices; they become our trusted confidants. With the appearance of short-range wireless network technology like Bluetooth (www.bluetooth.com), our personal mobile devices gain the ability to assist us in our daily social encounters. This is the world of wearable communities (Kortuem, Segall, & Thompson, 1999a; Kortuem, Schneider, Suruda, Fickas, & Segall, 1999b).

Wearable computing pursues an interface ideal in which the computer persists and provides constant access to information services, senses and models context, augments and mediates the user's interactions with the environment, and interacts seamlessly with the user. But perception on the body is a relatively new endeavor, since appropriate sensors are just now becoming available. While much study has centered on low-attention interfaces for automobiles and aircraft, little has been done for users of personal head-up displays. Furthermore, as with any wireless mobile device, the amount of power and the type of services available can constrain networking. Another serious issue is open standards to enable interoperability between different services. For example, only one long-range radio should be necessary to provide telephony, text messaging, Global Positioning System (GPS) correction signals, and so on.

Apart from wearable communities, ongoing research on 3D virtual reality (Huang, Eliens, & Visser, 2002) and online communities has shown that a new, real-time, multimedia community can be created and that it would not be long that users would be able to explore virtual worlds that resemble most closely their off-line counterparts. As a number of specific technologies mature, the expressiveness and functionality that computer-mediated communication within such communities require will advance to the next level.

In virtual worlds, people could have their own personal "avatars", controlling their moves by pointing their mouse, pressing keyboard keys, or using simple voice commands (Shen, Radakrishnan, &

Georganas, 2002).Utilizing advanced interfaces, people could gather and interact in public and private spaces, own and share objects, and spend lots of time online (Adler & Christopher, 2003). Interactive characters in virtual worlds will play supportive and helpful roles interacting with their users or other members of the community through natural forms of conversation and gesture, keeping at the same time track of relationships and preferences in a personalized database which they will be updating constantly (Elliott & Brzezinski, 1998). This functionality can be accomplished with several artificial intelligence techniques, but arguably the most viable means for achieving it is via the application of software agents. Through their learning, autonomy, cooperation, and flexibility capabilities, software agents hold the potential and will eventually become a significant part of every virtual world of 3D representations with which agents can examine, interact, and use.

Although virtual worlds seem to be the way forward, there are many issues related to their implementation and usability that must be resolved for this to be achieved (Malhotra, Gosain, & Hars, 1997; Stolterman, 1999). For example, the graphical part of such software consists of several megabytes of data that a user must download. Even if the software is downloaded from the server to the user's client, slow or even medium Internet connection speeds will be prohibitive for the use of the online, real time nature of such applications and the demands they make. Furthermore, as it has been observed, the growth of the use of software agents in virtual worlds have been moderate in comparison to other applications mainly because of the inherent complexities in developing such artifacts required to function within the large ontological spaces of virtual worlds.

For these and other reasons, virtual worlds are still in their infancy. Two of the most advanced examples of online virtual worlds line are "The Palace" from Electric Communities, and "Worlds Away" from Fujitsu. They both provide some form of avatars, but only for chat purposes, with low graphics and without the appearance of agents. As soon as researchers and developers overcome the various obstacles, virtual worlds will achieve the popularity and growth needed and safeguard thus their place in the evolutionary path of virtual communities.

CONCLUSION

It must be said that virtual communities, although they rely on a variety of technologies, are not about technology. They are about people (Churchill & Bly, 2000). They represent a new kind of social institution that provides new ways for individuals with common interests to meet and interact with one another. Because of this they also represent an important new economic force that is opening up new avenues of interaction between companies and consumers.

However, because of the complexities emanating from their socio-technical nature, building and sustaining a virtual (or indeed a physical) community, is not a simple matter. As Gerry McGovern of Digital Places (Digital Places, 1998) has observed, "communities are complex, difficult things. They take time to grow and are slow to change. What makes them strong can make them scary to the outsider or to the member who wishes to be different. Those who seek to work with 'online communities' need to understand that they will not be easily packaged into three-year business plans."

REFERENCES

Adler, R.P. & Christopher, A.S. (2003). Internet community primer: Overview and business opportunities. Last accessed November 15, 2003, at *http://www.digiplaces.com/pages/primer_00_toc.html*

Armstrong, A. & Hagel, B. (1997). *Net gain: Expanding markets through virtual communities*. Harvard Business School Press.

Churchill, E.F. & Bly, S. (2000). Culture vultures: Considering culture and communication in virtual environments. *ACM SIGGROUP Bulletin, 21*(1), 6-11.

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities.* University of Chicago Press.

Cronin, K. (1995). Virtual communities: A brief history. Last accessed October 24, 2003, at http://*www.ucalgary.ca/~dabrent/380/webproj/kathleen.html*

Digital Places (1998). Virtual communities. Last accessed November 15, 2003, at *http://www.digiplaces.com/pages/primer_01.html*

Elliott, C. & Brzezinski J. (1998). Autonomous agents as synthetic characters. *AI Magazine, 19*(2), 13–30.

Frohberg, D. (2003). Communities - The MOBIlearn perspective. Last accessed May 23, 2004, at *www.ifi.unizh.ch/ im/imgt/fileadmin/publications/frohberg_Communities_-_The_MOBIlearn_perspective.pdf*

Huang, Z., Eliens, A., & Visser, C. (2002). 3d agent-based virtual communities. *Proceedings of the Seventh International Conference on 3D Web Technology*, Tempe, AZ., 137-143.

Igbaria, M. (1999). The driving forces in the virtual society. *Communications of the ACM, 42*(12), 64-70.

Johnston, E., Raizada, L. & Cronin, K. (1996). Virtual communities. Last accessed October 25, 2003, at *http://www.ucalgary.ca/~dabrent/380/webproj/vircomm2.html*

Kortuem, G., Schneider, J., Suruda, J., Fickas, S., & Segall, Z. (1999b). When cyborgs meet: Building communities of cooperating wearable agents. *Proceedings of the 3rd International Symposium on Wearable Computers*, San Francisco.

Kortuem, G., Segall, Z., & Thompson, T.G.C. (1999a). Close encounters: Supporting mobile collaboration through interchange of user profiles. *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing*, Karlsruhe, Germany.

Köth, C. (2003). Virtual communities and the public library. Last accessed November 13, 2003, at *http://www.slais.ubc.ca/courses/libr500/02-03-wt2/www/C_Koth/introduction.htm*

Malhotra, A., Gosain, S., & Hars, A. (1997). Evolution of a virtual community: Understanding design issues through a longitudinal study. *Proceedings of the Eighteenth International Conference on Information Systems*, Atlanta, GA., 59-74.

Roberts, T. (1998). Are newsgroups virtual communities? *Proceedings of the SIGCHI conference on Human factors in Computing Systems*, Los Angeles, CA., 360-367.

Schubert, P. (1999). *Aufbau und management virtueller geschoftsgemeinschaften in electronic commerce.* Umgebungen, University of St.Gallen. Unpublished Dissertation.

Schubert, P. & Ginsburg, M. (2000). Virtual communities of transaction: The role of personalization in electronic commerce. *Electronics Markets Journal, 10*(1), 45-55.

Shen, X., Radakrishnan, T., & Georganas, N. (2002). vCOM: Electronic commerce in a collaborative virtual world. *Electronic Commerce Research and Applications, 1*, 281-300.

Stolterman, E. (1999). Technology matters in virtual communities. *ACM SIGGROUP Bulletin*, *20*(2), 7-9.

KEY TERMS

AOL: A US online service provider based in Vienna, Virginia, USA. AOL claims to be the largest and fastest growing provider of online services in the world, with the most active subscriber base. AOL offers its three million subscribers electronic mail, interactive newspapers and magazines, conferencing, software libraries, computing support, and online classes amongst other services.

ARPAnet: The first multisite, packet-switched network. ARPAnet was designed to support the Advanced Research Projects Agency (ARPA) for the transferring of files and resource sharing.

FTP: File Transfer Protocol. Protocol that allows users to copy files between their local system and any system they can reach on the network.

GPS: Global Positioning System. GPS was developed by the US DOD to allow the military to accurately determine their precise location anywhere in the world. GPS uses a collection of 24 satellites positioned in orbit to allow a person who has the proper equipment to automatically have their position triangulated to determine their location.

PDA: Personal Digital Assistant. A PDA is a small digital device that is used to store information such as phone numbers, addresses, schedules, calendars, and so on. A PDA may also be referred to as a handheld device or as a Palm. The Palm Pilot was one of the original PDAs and is now joined by others such as Palm Tungsten, HP IPaq, Palm Zire, and the Toshiba Pocket PC.

Telnet: A terminal emulation program for TCP/IP networks such as the Internet. The Telnet program runs on the client computer and connects it to a host on the network. Commands can be then entered through the Telnet program and they will be executed as if the user was entering them directly on the server console.

USENET: Usenet is a worldwide distributed discussion system. It consists of a set of "newsgroups" with names that are classified hierarchically by subject. "Articles" or "messages" are "posted" to these newsgroups by people on computers with the appropriate software—these articles are then broadcast to other interconnected computer systems via a wide variety of networks. Some newsgroups are "moderated"; in these newsgroups, the articles are first sent to a moderator for approval before appearing in the newsgroup. Usenet is available on a wide variety of computer systems and networks, but the bulk of modern Usenet traffic is transported over either the Internet or UUCP.

UUCP: Unix to Unix Copy. A Unix utility program and protocol that allows one Unix system to send files to another via a serial line which may be a cable going directly from one machine's serial port to another's or may involve a modem at each end of a telephone line.

Virtual Communities on the Internet

Abhijit Roy
Loyola College in Maryland, USA

INTRODUCTION

With the advent of the Internet a little over a decade ago, technology has enabled communities to move beyond the physical face-to-face contacts to the virtual realm of the World Wide Web. With the advent of highways in the 1950s and 1960s, communities were created in suburbia. The Internet, on the other hand, over the last fifteen years, has enabled the creation of a myriad of virtual communities that have limitless boundaries around the entire globe.

This paper begins by providing a definition of the term *virtual communities* and then describing several typologies of this phenomenon. The various motivations for joining communities, how marketers create social bonds that enhance social relationships, as well as strategies used by firms in building virtual communities also are discussed. We conclude by discussing strategies for managing virtual communities, researching them, as well as directions for future research.

DEFINITION

A community refers to an evolving group of people communicating and acting together to reach a common goal. It creates a sense of membership through involvement or shared common interests. It has been considered a closed system with relatively stable membership, which demonstrates little or no connection to other communities (Anderson 1999).

With the rapid growth of the Internet, the geographic boundaries constraining the limits of communities are no longer a factor, and the functions of maintaining a community can be fulfilled virtually from anywhere on the globe. This is the basic essence of a virtual community. Several authors have attempted to provide a formal definition of the term for semantic clarifications. The major definitions are as follows:

Social aggregations that emerge from the Net when
enough people carry on public discussions long enough , with sufficient human feeling, to form webs of personal relationships in cyberspace *(Rheingold, 1993, p. 5).*

Groups of people who communicate with each other via electronic media, rather than face-to-face *(Romm, Pliskin, & Clarke, 1997, p. 1997).*

Computer mediated spaces where there is a potential for an integration of content and communication with an emphasis on member generated content (Hagel & Armstrong, 1997, p. 134).

Virtual Publics are symbolically delineated com puter mediated spaces, whose existence is relatively transparent and open, that allow groups of individu als to attend and contribute to a similar set of computer-mediated interpersonal interactions *(Jones & Rafaeli, 2000, p. 215).*

While Rheingold (1993) provides one of the earliest definitions of the term, and one that is most quoted in the literature (Kozinets, 1999), many may question whether "with sufficient human feeling" is a necessary conditions for virtual community formation. Romm, Pliskin, and Clark's (1997) definition may not sufficiently distinguish it from general Web sites. Hagel and Armstrong (1997) emphasize member-generated content, while Jones and Rafaeli (2000) use the term *virtual publics* instead of virtual community. Based on these definitions, the term may be simply defined as follows:

A group of individuals with common interests who interact with one another on the Internet.

Typologies of Virtual Communities

Virtual communities come in different shapes and

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

sizes and may have memberships of a few dozen to millions of individuals. These communities may extend from active forums like discussion groups and chat rooms to passive ones like e-mails and bulletin boards. Given that these communities are not geographically constrained, their size can be much bigger than typical physical communities,and many millions of them exist on the Internet. Uncovering archetype or gestalt patterns is fundamental to the study of social science and research, and several authors have proposed classification schemes for configurations of virtual communities.

Lee, Vogel, and Limayem (2003), in their review of classification schemes of virtual communities, identify Hagel and Armstrong's (1997) and Jones and Rafaeli's (2000) typologies as being the most popularly referenced. Kozinets (1999) also delineates four kinds of virtual communities. These three typologies are reviewed, and a further popular typology of affinity groups proposed by Macchiette and Roy (1992), as applied to the virtual environment, is also proposed.

Hagel and Armstrong (1997) propose four major types of virtual communities based on people's desire to meet basic human needs: interest, relationship, fantasy and transaction. Jones and Rafaeli (2000) further segment these communities by social structure (i.e., communities formed based on social networks, such as virtual voluntary associations, cyber inns, etc.) and technology base (i.e., types of technology platforms, such as e-mail lists, Usenet groups, etc.).

Kozinets (1999) proposed the four types of communities as dungeons (i.e., virtual environments where players interact, such as for online video games); circles (i.e., interest-structured collection of common interests); rooms (i.e., computer-mediated environments where people interact socially in real time); and boards (i.e., virtual communities organized around interest-specific bulletin boards).

Finally, Macchiette and Roy (1992) proposed a typology of affinity communities that also can be used for classifying virtual communities. They defined communities as either being: professional (e.g., doctors, lawyers, etc.), common interest (e.g., hobbies, interests), demographic (e.g., by gender, age, etc.), cause-based (e.g., Sierra Club, Green Peace), and marketer generated (e.g., Disney, Nintendo) communities. These communities also may be constructed in the virtual environment.

It is also interesting to make other dichotomous distinctions of virtual communities such as (a) formal (e.g., associations) vs. informal communities; (b) commercial (offers goods and services to make revenues that, in turn, fuel community operations) vs. noncommercial (communities created from the ground up by a group of individuals, such as one interested in stamp collection); and (c) open or public (where everyone, regardless of their qualifications and individual profile, can enter the community and participate) vs. closed or private (where outsiders are not allowed into the community, or where membership is very difficult to obtain).

Virtual Communities: Motivations, Mode of Participation, Characteristics, and Benefits

Rayport and Jaworski (2004) present a model of how the various components of a virtual community can be integrated. An adapted version of the model is shown in Figure 1. The model illustrates how members' motivations for joining the virtual community, their mode of participation, and the community's degree of connectedness, in many ways determine the characteristics of the community, which, in turn, influence the benefits sought by the members in these communities. The various components of the model are discussed next.

Motivations

A member's reasons for joining a community may depend on a wide range of factors, such as affiliation (others like them are members of the community), information (about experiences, ideas, and issues), recreation (meeting people, playing around, sharing stories, etc.), or transaction (e.g., those who join a Web site for buying and trading possessions).

Mode of Participation

Participation can occur in a myriad of ways (e.g., through e-mails, chat rooms, discussion groups, online events, bulletin boards, etc.). Some (i.e., discussion groups, chat rooms) have more active members than passive members (e.g., e-mail or bulletin board).

Characteristics of Virtual Communities

With the growth and maturity of virtual communities, certain characteristics are prevalent. Adler and Christopher (1999) identify six such characteristics:

- **Cohesion:** Members seek a sense of belonging and develop group identity over time.
- **Relationships:** Community members interact and develop friendships over time.
- **Effectiveness:** The group has an impact on members' lives.
- **Help:** Community members feel comfortable asking and receiving help from each other.
- **Language:** Members develop shared communication tools that have a unique meaning within the community.
- **Self-Regulation:** The community develops a system for policing itself and sets ground rules of operation.

Benefits to Members

Adler and Christopher (1999) further point out that the members of the virtual community develop various emotional benefits, depending on the communities that they join. They include: inclusion, shared information and experiences, need fulfillment, and mutual influence, among others.

Degree of Connectedness in Virtual Communities

The degree of connectedness in virtual communities also plays a significant role in how a virtual community develops. They can be classified as weak, limited, or strong. This primarily depends on the degree of interactivity between and among members.

- **Weak:** Members of these sites have no opportunities to interact with each other on a one-on-one basis (e.g., newspaper Web sites, corporate Web sites, etc.).
- **Limited:** These communities offer limited opportunities for members to interact with other (e.g., reading, posting information or opinions, etc.).
- **Strong:** These communities offer chat rooms and message boards and allow users to form strong bonds with each other.

Research has shown that both strong and weak connectednesses have their own advantages. While weak ties are shown to facilitate such tasks as finding jobs (Granovetter, 1973), strong ties are required to facilitate major changes in the communities (Krackhardt, 1992) (see Figure 1).

Stages of Virtual Community Life Cycle

Kim (2000) proposes a five-stage virtual community building process that progresses as follows:

1. **Visitors:** These are individuals who lurk in the virtual community, yet don't participate in them.
2. **Novices:** They are new members or "newbies" who are usually passive and are busy learning the rules and culture of the virtual community and, thus, are not actively engaged in it.
3. **Regulars:** They are established members comfortably participating in the exchanges and make up the largest segment of the virtual community.
4. **Leaders:** These members are volunteers, contractors, and staff who create topics and plan activities that keep the virtual community running.
5. **Elders:** They are respected members of the virtual community, who are always eager to share their knowledge and pass along the culture of the community to the newer members.

Mohammed et al. (2004) further suggest four relationship stages: awareness, exploration/expansion, commitment and dissolution, and the varying level of intensity patterns as virtual community members go through membership life cycle. At the initial awareness stage, members have the lowest intensity levels and are likely to be considered visitors up until the exploration stage. At this second stage, these novices develop greater intensity and commitment to the site. The equity building efforts over time translate into the virtual members becoming regulars and, subsequently, leaders or elders. Finally, over

Figure 1. Virtual communities: Motivations, mode of participation, characteristics, and benefits

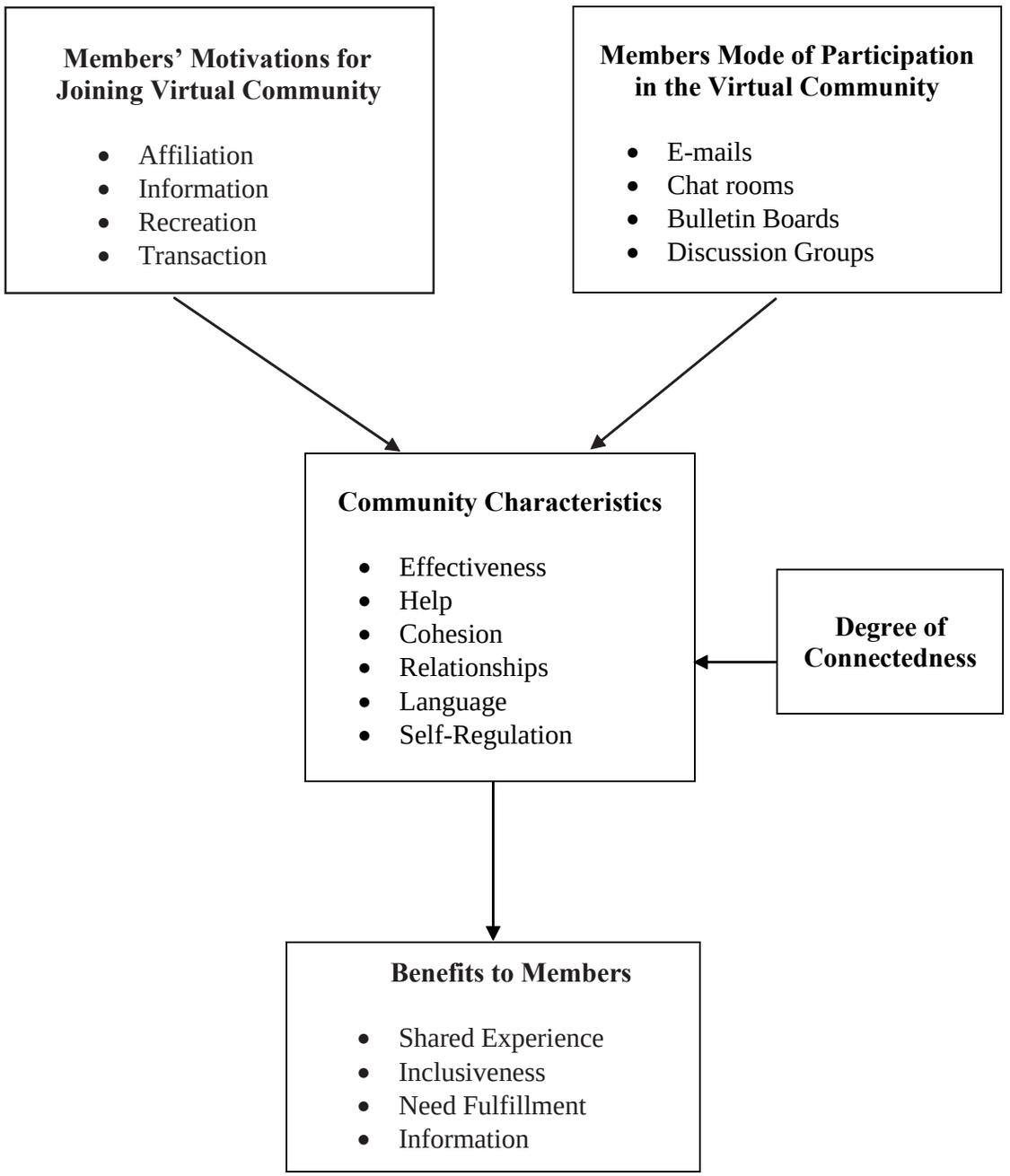
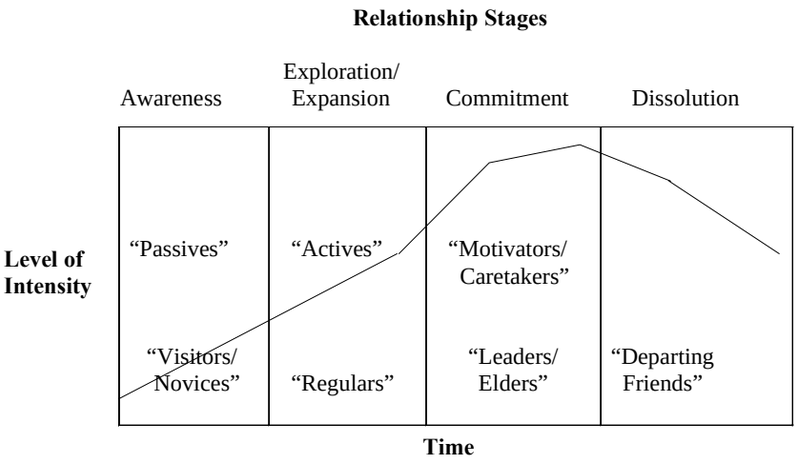


Figure 2: Intensity patterns of the different types of virtual communities at various relationship stages (Adapted from Mohammed et al., 2004)



time, even the most committed members outgrow a community and become departing friends. Figure 2 illustrates these stages.

Farmer (1994) had earlier described four similar stages through which individuals in virtual communities mature. According to him, members begin as *passives* (attending a community, yet not actively engaging in it), and then go on to become *actives* (participating in communities and taking part in conversations). The highest levels of participation are displayed by *motivators* (those who create conversation topics and plan activities) and *caretakers* (those who act as intermediaries between members).

The passives are analogous to the visitors and novices; the actives are similar to the regulars, while the motivators and caretakers are equivalent to the leaders and elders in the Mohammed et al. (2004) model.

Strategies for Managing Successful Virtual Communities

Duffy (1999) outlines the eight critical factors for community success, as recommended by Accenture, the Management Company. They are:

- Increasing traffic and participation in the community.
- Focusing on the needs of the members by using facilitators and coordinators.
- Keeping the interest high by provoking controversial issues.
- Involving the community members in activities and recruiting.
- Providing tools and activities for member use.
- Managing the cultural environment.
- Encouraging free sharing of opinions and information.
- Obtaining financial sponsorship.

Researching Virtual Communities

Kozinets (2002) suggests using netnography, involving ethnographic techniques in studying virtual communities for providing insights into the symbolism, meanings, and consumption patterns of virtual communities. The method is derived from ethnography, which was developed in the field of anthropology. Netnography involves the study of distinctive meanings, practices, and artifacts of virtual communities.

Rather than approaching the problem from a positivistic or scientific point of view, where a researcher begins with a theory, develops and tests hypotheses, and draws conclusions, netnography approaches the construction of meaning in virtual communities in an open-ended manner using inductive techniques and grounded theory. Since the research technique by nature is unobtrusive, ethical research guidelines must strictly be followed, such as (a) fully disclosing his or her presence, affiliations,

and intentions to virtual community members; (b) ensuring confidentiality and anonymity to respondents, and (c) seeking and incorporating feedback from the online community being researched.

FUTURE RESEARCH ISSUES

There are several issues relating to virtual communities that are worth investigating. First and foremost is the issue of whether or not they facilitate socialization or whether they are a threat to civilization. Some see them as a way of enhancing social capital between families, friends, and acquaintances, empowering individuals and organizations, creating new ways of relating to each other. Innovative firms leverage this power to create growth and create loyal customers. Others see them as a far cry from the regular face-to-face interactions, creating weak ties between strangers instead of strengthening existing ties between friends and neighbors.

Other issues deal with how to integrate online and off-line communities and how to develop appropriate metrics for such integration. How can these communities reduce member churn and build loyalty? What are the appropriate metrics for measuring community strength? Hanson (2000) suggests using content attractiveness, member loyalty, member profiles, and transaction offerings as possible metrics for measuring this phenomenon. Under what circumstances is loyalty developed through member-to-member relationships vs. content attractiveness vs. the transaction offerings? What is the most appropriate way to classify the typologies and taxonomies of these communities? How are intentional social actions generated in such communities? (Bagozzi and Dholakia 2002). Are virtual communities likely to replace regular face-to-face associations in the long run?

Virtual communities of all shapes and forms are rapidly evolving and creating values for their respective members. Many such communities have millions of members. These communities will continue to attract the interest of researchers from a wide range of academic fields in the future.

REFERENCES

Adler, R.P., & Christopher, A.J. (1999). Virtual Communities. In C.F. Haylock & L. Muscarella (Ed.), *Net success,* Holbrook, MA: Adams Media, pp. 36-59.

Anderson, W.T. (1999). Communities in a world of open systems. *Futures, 31*, 457-463.

Bagozzi, R.P., & Dholakia, U.M. (2002). Intentional social action in virtual cCommunities. *Journal of Interactive Marketing*, *16*(2), 2-21.

Cindio, F.D., Gentile, O., Grew, P., & Redolfi, D. (2003). Community networks: Rules of behavior and social structure. *Information Society*, *19*(5), 395-404.

Duffy, D. (1999, October 25). It takes an e-village. *CIO Magazine*. Retrieved from *http://www.cio.com/archive/enterprise/101599_virtent.html*

Farmer, F.R. (1994). Social dimensions of habitat's citizenry. In C. Loeffler, & T. Anderson (Eds.), *The virtual reality* (pp. 87-95).New York, NY: Van Nostrand Reinhold.

Granovetter, M.S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*, 1360-1380.

Hagel, J., & Armstrong, A. (1997). *Net gain: Expanding markets through virtual communities.* Boston, MA: Harvard Business Press.

Hanson, W. (2000). *Principles of Internet marketing*. Cincinnati, OH: South-Western.

Igbaria, M. (1999). The driving forces in the virtual society. *Association for Computing Machinery: Communications of the ACM, 42*(12), 64-70.

Jones, Q., & Rafeli, S. (2000). Time to split virtually: "Discourse architecture" and community building as means to creating vibrant virtual metropolises. *International Journal of Electronic Commerce and Business Media*, *10*(4), 214-223.

Kim, A.J. (2000). *Community building on the Web*. Berkeley, CA: Peachpit Press.

Kozinets, R.V. (1999). E-tribalized marketing? The strategic implications of virtual communities of con-

sumption. *European Management Journal, 17*(3), 252-264.

Kozinets, R.V. (2002). The field behind the screen: Using the method of netnography to research market-oriented virtual communities. *Journal of Marketing Research, 39*, 61-72.

Krackhardt, D. (1992). The strength of strong ties: The importance of philos in organizations. In N. Nohria, & R. Eccles (Eds.), *Networks and organizations: Structure, firm and action*. Boston, MA: Harvard Business Press.

Lee, F.S.L., Vogel D., & Limayem, D. (2003). Virtual community informatics: A review and research agenda. *Journal of Information Technology Theory and Application, 5*(1), 47-61.

Luo, X. (2002). Trust production and privacy concerns on the Internet: A framework based on relationship marketing and social exchange theory. *Industrial Marketing Management, 31*(2), 111-118.

Macchiette, B., & Roy, A. (1992, Summer). Affinity marketing: What is it and how does it work? *Journal of Services Marketing*, 47-57.

Maclaran, P., & Catterall, M. (2002). Researching the social Web: Marketing information from virtual communities. *Marketing Intelligence and Planning, 20*(6), 319-326.

McWilliam, G. (2000, Spring). Building stronger brands through online communities. *Sloan Management Review*, 43-54.

Mohammed, R.A., Fisher, R., Jaworski, B.M., & Paddison, G.J. (2004). *Internet marketing: Building advantage in a networked economy*. New York: McGraw-Hill/Irwin.

Rayport, J.F., & Jaworski, B.J. (2004). *Introduction to e-commerce*. New York: McGraw-Hill.

Rheingold, H. (1993). *Virtual community: Homesteading on the electronic frontier*. Reading, MA: Addison Wesley.

Romm, C., Pliskin, N., & Clarke, R. (1997). Virtual communities and society: Toward an integrative three phase model. *International Journal of Information Management, 17*(4), 261-270.

Wellman, B., et al. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology, 22*, 213-238.

KEY TERMS

Affinity Communities: Communities that are based on profession, common interest, cause, demographic or marketer generated phenomenon.

Characteristics of Virtual Communities: Virtual communities are characterized by their level of cohesion, effectiveness, helpfulness of members, quality of the relationships, language, and self-regulatory mechanisms.

Netnography: Using ethnographic techniques to study virtual communities.

Stages of the Virtual Community Life Cycle: Virtual community members go through four relationship stages (i.e., awareness, exploration/expansion, commitment, and dissolution).

Virtual Community: A group of individuals with common interests, who interact with one another on the Internet.

Virtual Knowledge Space and Learning

Meliha Handzic
Sarajevo School of Science and Technology, BiH, Croatia

Joanne Chia Yi Lin
The University of New South Wales, Australia

INTRODUCTION

The growing importance of knowledge and innovations for modern organisations (Davenport, DeLong, & Breers, 1998; Drucker, 1998; Nonaka, 1998; Stewart, 1997), and increasing demands for new skills and capabilities suggest the need for improvement in the learning of future professional and managerial workers. This, in turn, requires an appropriate response from the education sector. So far, these demands have not been adequately addressed by management education (Seufert & Seufert, 1999). There are calls to base the learning more in reality, to make the learning and thought process visible in order to develop the learners' metacognition (Joyce & Weil, 1986), and to achieve better balance between the imparting of knowledge to the learner and the learner's own construction of it. It is also suggested that education should better nurture students' qualities such as problem solving, decision making, and creativity through self-directed as well as collaborative creativity and learning. These are skills that students will require in order to be successful in their future roles as innovative professionals and business people.

Given the crucial importance of knowledge and innovation for success in the knowledge economy, the main purpose of this study is to address the issue of students' learning in the context of graduate information-systems education. In particular, the article will investigate students' idea-generation behaviour and propose a Web-based knowledge space or k-space as a flexible learning tool to support their individual styles.

LITERATURE REVIEW ON IDEA GENERATION AND KNOWLEDGE MANAGEMENT

Idea generation can be defined as the production of novel and appropriate ideas, solutions, and work processes. A holistic view of idea generation has been recently provided by Shneiderman (2000). He differentiates three approaches: *inspirationalist*, which concentrates on the intuitive aspects of idea generation; *structuralist*, which emphasises the importance of previous work and methods in exploring different possible solutions; and *situationalist*, which focuses on the social context as a key part of the idea-generation process. Vandenbosch, Fay, & Saatciglu (2001) argue that most theories study idea generation in terms of individual characteristics, contexts in which ideas flourish, or details about processes in which ideas are developed. They argue that in real life, people do not come up with ideas in isolation, and that different combinations of contextual and personal characteristics may result in different but equally effective processes.

Taking a broader view of the cognitive-styles perspective, these researchers investigate the interrelatedness of idea generation, problem solving, and inquiry to explore the notion of archetype based on Churchman's (1971) system of inquiry. Figure 1 shows Vandenbosch et al.'s (2001) proposed classification scheme, which involves five idea-generation archetypes (Leibniz, Locke, Kant, Hegel, and Singer) based on individual approaches to information acquisition, change, relationships to others, and problem solving.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

Figure 1. Five idea-generation archetypes

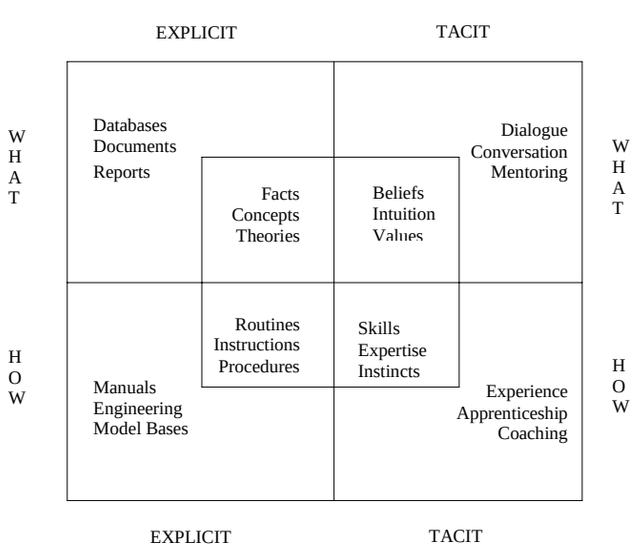
	Leibniz	Locke	Kant	Hegel	Singer
Information Acquisition	Searching	Searching	Scanning	Scanning	Scanning
Approach to Change	Maintaining	Reacting	Initiating	Initiating	Initiating
Relationship to Others	Directing	Mediating	Collaborating	Internalizing	Unpredictable
Problem Solving	Retaining	Converging	Diverging	Debating	Unpredictable

In summary, Leibnizian inquirers are seen primarily as incrementalists, placing a great deal of importance on what they already know. Lockean inquirers are known as the consensus builders, typically asking others to generate ideas and focusing on agreement. Kantians are viewed as searchers who combine ideas from diverse sources and unusual associations. Hegelians are known as debaters arguing internally with themselves to develop ideas. Finally, Singerians are considered the most flexible inquirers, comfortable with and employing all systems of inquiry.

Knowledge management (KM) is the most recent response to the need to better understand and manage knowledge for success or survival. Knowledge management is concerned with the processes of creation, acquisition, organisation, and transfer of knowledge, as well as organisational, cultural, technological, and measurement enablers that may facilitate these processes and foster the development of working knowledge and performance. The central task of those concerned with knowledge management is to determine the best ways to cultivate, nurture, and exploit knowledge at individual and organisational levels. In other words, knowledge management needs to ensure that the right knowledge gets to the right people at the right time, and it helps people share and put knowledge into action in ways that strive to improve organisational performance (O'Dell & Grayson, 1998).

Different approaches for knowledge management are possible, originating from cognitive psychology, philosophy, education, science, finance, and information technology. While all are valuable, they deliver only a partial view on the whole topic. Davenport and Prusak (1998) maintain that it is only possible to

Figure2. Core knowledge-management framework



realise the full power of knowledge by taking a holistic, ecological approach to knowledge management. Following the holistic-approach recommendations by Davenport and Prusak, a core model of KM presented in Figure 2 was proposed by Handzic and Jamieson (2001) for conducting research in KM.

Given the current infancy of knowledge-management theory and practice, there is little research done regarding the value of knowledge-management initiatives in idea generation. Therefore, the first objective of this research is to conduct an empirical study to categorise students' idea-generation behaviour using the Vandenbosch et al. (2001) classification as a guideline. The second objective is to apply the knowledge-management principles described above to formulate a suitable design for a Web-based k-space that would be an innovative learning tool able to accommodate the idea-generation styles of different students.

If it is clear how academic students approach knowledge, it will be easier to provide better support systems to encourage knowledge sharing and creativity. This can be helpful for both academic staff and the students. Teachers can assist students more readily if they understand how the students learn and acquire information. Students can develop better learning skills and understand the importance of managing their knowledge as an asset for their future career paths.

SURVEY OF STUDENTS' IDEA-GENERATION STYLES

A survey study was conducted to determine students' idea-generation behaviour based on the work by Vandenbosch et al. (2001). Subjects for this study were 72 students enrolled in the Knowledge Management Systems and Technology course at the University of New South Wales (UNSW). The participating students came from different academic and cultural backgrounds, providing a diverse sampling group. The researchers chose to focus on students' idea-generation behaviour as it was felt that students were "knowledge workers" with much potential to influence their future career organisations with their preferred style.

With respect to information acquisition, respondents were assessed in terms of their inclination to *search* for information with a predetermined agenda and focus, or *scan* for information with a broad agenda and view. With regard to their approach to change, the participants were assessed by their likelihood to *react* to a real or potential problem, *maintain* the status quo, or *initiate* change in order to increase their capacity to influence the environment. Respondents' relationships to others were assessed in terms of their tendency to *direct* others, *mediate* by empowering others, *collaborate* by joining others in generating ideas and making decisions, or *internalise* by attempting to resolve problems individually by playing through multiple scenarios. Finally, subjects' problem-solving methods were assessed in terms of their tendency to *retain* information by focusing on ideas that complement and affirm their own approach, *converge* by closing in upon ideas to find an agreeable solution, *diverge* by considering and expanding upon many ideas to develop a specific solution, or *debate* by having dialectic discussions of several ideas in order to develop a specific solution.

From the responses obtained for the percentage of time spent on different inquiry methods, we were able to categorise the participants into the five archetypes. The participants clearly reflected four different types. The results revealed the following composition across these types: Leibnizian (57%), Lockean (29%), Kantian (8%), and Hegelian (6%). These results indicate a dominant Leibnitz type, followed by Lockean, Kantian, and Hegelian. An interesting result is that no Singerian- (0%) type representatives were found in the group. Singerians are assumed to be very flexible and adaptive, and would adopt a variety of different methods for study. However, subjects in this research opted for a dominant method.

The main findings of this study provide strong support for the extended cognitive-style perspective on learning that interrelates idea generation, problem solving, and inquiry approaches. The current study discovered significant differences among student participants in their systemic approaches to knowledge based on Churchman's (1971) interpretation of philosophies. As such, these findings have a number of important implications for education. More specifically, they suggest the need to design an appropriate learning-teaching support system to respond to different, individual students' needs. In

addition, learning-teaching support systems need to address the needs of the class as a whole, which will generally be a combination of different archetypes.

First, Leibnizian tendencies show that over one half of students believe that knowledge should come from study. They place a great emphasis on facts and they learn using formal logic to make inferences of causes and effects. The Leibnizian type tends to rely heavily on theory and what he or she has already learned. Therefore, providing instruction together with lecture notes and textbooks would benefit Leibnizians' learning style. A typical Leibnizian response to acquiring information is reading from documents and reports provided.

About one third of the Lockeans employ experimental and consensual reasoning. They learn by observing the world, sharing their observations, and creating consensus about what has been observed. To benefit this learning style, group discussions and group assignments could assist students to increase their learning interest and performance, and also to provide feedback. These are all important sources that help the group to reach consensus. Experimental results of an issue-identification study conducted by Aurum, Handzic, and Gardiner (2002) indicate that groups of four to five students are optimal.

A somewhat smaller percentage of students were identified as Kantians, also known as searchers. They like to seek knowledge from diverse sources and are very broad in their search. They tend to scan internal and external environments for purposeful knowledge. Therefore, in order to satisfy Kantians' needs in learning, educators should provide them with many rich resources and readings outside the textbook.

There were also some Hegelians, the debaters who tend to synthesise opposing models into a new worldview. They rely upon the dialectic to resolve opposing viewpoints, and they may generate an entirely new idea as a result. Therefore, facilitating the dialectic process would enable the Hegelian to acquire knowledge and encourage him or her to generate innovative ideas.

Although our survey uncovered none in this group, Singerian inquirers are Renaissance people who employ all systems of inquiry. They constantly question and work very hard. In other words, their learning styles are very flexible. Therefore, providing potential Singerians with all the above features and facilities enables them to gain knowledge and generate new insights in their studies.

VIRTUAL-KNOWLEDGE-SPACE DESIGN PHILOSOPHY AND FEATURES

Designing an appropriate learning-teaching support system to respond to different, individual students' needs in one class presents a major challenge. However, new developments in information and communication technologies such as the Internet and Word Wide Web may be able to provide a valuable technology base for innovative learning tools. A Web-based k-space designed on the principles of knowledge management may provide the necessary place to exchange, share, capture, discover, and obtain knowledge resources for a learner. It also may be a valuable virtual learning community for educators and students alike to share and discuss matters relating to the course. In summary, it can enable or facilitate knowledge processes and enhance learning performance. The k-space described in the following section is a good example of one such learning tool that can support the needs of different students.

A conceptual model of the k-space presented in Figure 3 was developed by mapping the core knowledge-management framework presented in Figure 2 (see previous section) to the graduate course learning-teaching context. The resulting model is a 2x2 matrix with Explicit and Tacit (knowledge) as columns, and Know What and Know How as rows. The Explicit-Know-What quadrant consists of the availability of information resources that can be used by students to acquire explicit knowledge. For example, all the students can obtain explicit knowledge found in lecture notes, relevant resources, and databases. The Explicit-Know-How quadrant contains search facilities, and rules and patterns discovered by individual students while finding information. The Tacit-Know-What section consists of areas such as discussion forums and announcements that enable students and educators to share information. Finally, the Tacit-Know-How quadrant supports the individual's tacit knowledge learning that can be gained through personal experience and activities such as assignments, tutorials, and so forth.

Figure 3. A conceptual model of a virtual k-space

Knowledge Types	Explicit	Tacit
Know What	Availability of information • Lecture notes • Relevant resources • Databases	Sharing of information • Discussion forums • Announcement • News
Know How	Finding of information • FAQs (frequently asked questions) • Classifying • Search facilities	Learning by doing • Exercises • Simulation games • Feedback/guidance

The model suggests the following.

- Students' course knowledge can be enhanced by enabling and facilitating the availability, sharing, and finding of relevant information, as well as learning by doing.
- In order to minimize information overload, it is also suggested by the framework that students should be supported by intelligent search and mining facilities.
- The framework also recognizes the importance of tacit knowledge. Past researchers show that students may benefit from sharing knowledge with others as well as from interaction with their peers (Handzic & Low, 2002; Handzic & Tolhurst, 2002).
- Self-directed learning such as assignments and self-paced online learning sessions with continuous guidance and feedback would respond to the need for the cultivation of students' skills for problem solving, decision making, as well as creativity.

The knowledge-management principles described above were used for the development of a graduate course k-space at UNSW. The design presented here is the work of students attending the Knowledge Management Systems and Technology course at UNSW (for more details, see Chong, Jonson, & Chan, 2001). The main objective of the k-space was to provide students with a one-stop point of interaction for all their study needs: a portal that students could go to and obtain lecture notes, assignments, reference materials, discussions, surveys, search facilities, links, and many other useful tools. The idea of each section in the k-space was to implement one or more facilities from a quadrant of the knowledge matrix. In addition, each section supported different idea-generation and learning styles. Our study found that different students have preferences for different methods, and if their preferred method is provided, the productivity and enjoyment of the course will be increased.

The specific k-space design presented here consists of several parts as follows:

- The Home section contains announcements, quotes, and news, as well as evaluation forms, polls with results, and search facilities.
- Contact Details incorporates a contact person, consultation hours, class and lab venues, additional help, and comments.
- Course Details includes recommended texts, lecture notes, recommended readings, assignments, and the course outline. It is expected to be of particular value to Leibnizians, who rely heavily on theory and typically acquire knowledge by studying documents and reports provided.
- Discussion Forum has assignment discussions, lecture comments, additional user posts, and search-posting tools. It represents an ideal support for Lockeans who like to share their observations and create consensus.

- The Resources section contains university links, research papers, field-related Web sites, and search engines. It provides an ideal learning support for Kantians, who typically seek knowledge by scanning a wide variety of external sources.
- Solution Finder provides questions and answers, as well as simulation games. It is envisaged as a learning space for Hegelians, who tend to construct and internally debate different viewpoints and generate new solutions.

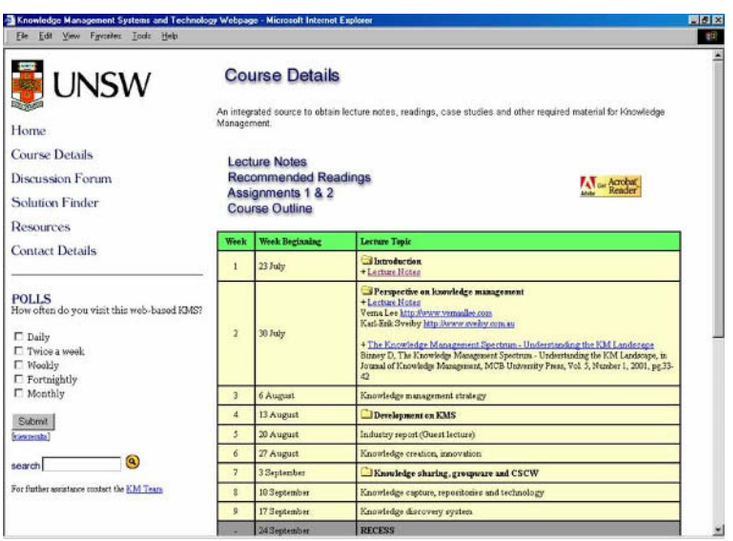
As an example, the Course Details section of the k-space can be seen in Figure 4. The purpose of the Course Details section is to allow users to obtain information necessary to assist them with the course. Course Details provides an overview of the course schedule over the 14 weeks. This page outlines the week number, week beginning date, and the lecture topics. The first two headings of the course outline are self-explanatory. Under the Lecture Topic heading, it contains folders for each of the 14 weeks. Furthermore, the Acrobat Reader software is available for download as all the files are saved in PDF (Portable Document Format) and users can easily gain access to these notes.

In summary, the k-space supports course learning by facilitating knowledge processes that foster the development of relevant knowledge for different types of students. It allows for taking small steps and incremental learning (Leibnizian), as well as building consensus (Lockean). More importantly, the k-site encourages higher forms of inquiry involving forming associations and combining information from diverse places (Kantian), and constructing ideas through internal debate of all the factors (Hegelian). Finally, it allows for flexibility by providing various forms of inquiry (Singerian) if students who use them all exist. Other benefits of the proposed k-space for students include enabling global access, ease of use, self-service, and collaboration. For educators, it provides a means to reduce paper work and to publish and maintain useful and dynamic information in a variety of forms.

FUTURE CHALLENGES FOR ONLINE EDUCATION AND TECHNOLOGY

Technology is currently driving a profound transformation of the learning industry. In response to the growing demand for education in the knowledge-based economy, universities and colleges are offering thousands of online courses, thus changing the traditional classroom-based methods of teaching and learning. Researchers and practitioners are predicting that the current trend will continue. However, while many institutions are developing and

Figure 4. Course details section of virtual k-space



using Web-based courses, little is known about their value in improving the quality of students' learning experiences. The underlying assumption is that technology can create conducive learning environments for students. However, empirical evidence to support this assumption is lacking.

The literature suggests that technology-mediated learning environments may improve students' achievement, their attitudes toward learning, and their evaluation of the learning experience. It also suggests that technology may help to increase teacher-student interaction and to make learning more student centred. In addition, proponents of virtual learning environments suggest that they can potentially eliminate geographic barriers while providing increased convenience, flexibility, currency of material, retention of students, individualised learning, and feedback over traditional classrooms. In contrast, other researchers suggest that technology-mediated learning environments may lead to student feelings of isolation, frustration, anxiety, and confusion. It may also result in reduced interest in the subject matter and questionable learner achievement (Piccoli, Ahmad, & Ives, 2001).

Given the growing interest in online education and the general lack of empirical studies examining the effectiveness of the technology-mediated learning environments, future research needs to address this issue by undertaking empirical research in different learning contexts and among different learners. It also needs to extend the current k-space development effort to include different and more advanced multimedia tools in order to find ways for providing a better and more satisfying learning experience for students of online courses.

CONCLUSION

The main objective of this article was to explore students' idea-generation behaviour based on Churchman's (1971) interpretation of different philosophies and to suggest a Web-based k-space to support different, individual approaches. The study has shown that Leibnizians make up the majority of the student population surveyed, followed by Lockeans, Kantians, and Hegelians. No Singerians were found at all in this study. These results suggest that different students would respond differently to alternative learning-teaching methods. Thus, it is important that each have access to methods that most stimulate his or her study interest, productivity, and enjoyment of the course.

A Web-based k-space designed on the principles of knowledge management is proposed as an innovative learning tool that can facilitate the learning needs of different students. The proposed virtual-space features provide support for the availability, finding, and sharing of information, as well as learning by doing. The main benefits for educators and students include reduction in paper-based documents and processes, ease of use, self-service for users, more readily available and dynamic information, global access, choice of media (documents may be read off a computer screen or printed), and greater support for an emerging electronically savvy culture. Future research is required to empirically test the implementation success of such a tool among various groups of students and in different course contexts.

REFERENCES

Aurum, A., Handzic, M., & Gardiner, A. (2002). Preparing IT professionals for the knowledge economy. *Proceedings of IRMA 2002*, Seattle, WA.

Chong, R., Jonson, C., & Chan, M. (2001). *Knowledge Website project report* [UNSW Knowledge Management Systems and Technology course assignment]. Sydney, Australia: University of New South Wales.

Churchman, C. W. (1971). *The design of inquiring systems: Basic concepts of systems and organisation.* New York: Basic Books, Inc.

Davenport, T. H., DeLong, D. W., & Breers, M. C. (1998, Winter). Successful knowledge management projects. *Sloan Management Review, 39*(2), 43-57.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge.* Boston: Harvard Business School Press.

Drucker, P. F. (1998). *The coming of the new organisation: Harvard business review on knowledge management.* Boston: Harvard Business School Press.

Handzic, M., & Jamieson, R. (2001). A knowledge management research framework for electronic commerce. *Proceedings of the IFIP TC8 Conference on E-Commerce/E-Business*, Salzburg, Austria.

Handzic, M., & Low, G. (2002). The impact of social interaction on performance of decision tasks of varying complexity. *OR Insight, 15*(1), 15-22.

Handzic, M., & Tolhurst, D. (2002). Evaluating interactive learning environment for decision making. *Educational Technology and Society, 5*(3), 113-122.

Joyce, B., & Weil, M. (1986). *Models of teaching.* Englewood Cliffs, NJ: Prentice-Hall.

Nonaka, I. (1998). *The knowledge-creating company: Harvard business review on knowledge management.* Boston: Harvard Business School Press.

O'Dell, C., & Grayson, C. J. (1998). *If only we knew what we know.* New York: Free Press.

Piccoli, G., Ahmad, R., & Ives, B. (2001). Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic IT skills training. *MIS Quarterly, 25*(4), 401-426.

Seufert, S., & Seufert, A. (1999). Collaborative learning environments for management education. *Proceedings of the 13th Annual Conference of the International Academy for Information Management*, 279-284.

Shneiderman, B. (2000, March). Creating creativity: User interfaces for supporting innovation. *ACM Transactions on Computer-Human Interaction, 7*(1), 114-138.

Stewart, T. A. (1997). *Intellectual capital: The new wealth of organisations.* New York: Doubleday.

Vandenbosch, B., Fay, S., & Saatciglu, A. (2001). *Where ideas come from: A systematic view of inquiry.* Sprouts: Working Papers on Information Environments, Systems and Organizations, 1(Fall), 1-23.

KEY TERMS

E-Learning (Distance Learning, Web-Based Learning, Online Learning): The application of computer and network technology in learning activities.

Home Page (Portal): A textual and graphical display that usually welcomes users to a Web site and provides a point of access to other static and dynamic Web pages.

Inquiry Style: An individual approach to learning based on dominant information acquisition, change, the relationship to others, and problem-solving behaviour.

Knowledge Management: A set of sociotechnological enablers and processes that move or modify knowledge resources and foster learning.

Virtual Knowledge Space (Web Site): A collection of Web pages that provides access to multiple sources of knowledge.

Web (World Wide Web, WWW): A system of universally accepted standards for storing, retrieving, formatting, and displaying knowledge in a networked environment.

Web Page: A hypermedia document that expresses the knowledge content in an artistic and dynamic fashion, combining text, graphics, audio, and video formats.

Virtual Learning Communities

Stewart T. Fleming
University of Otago, New Zealand

INTRODUCTION

The synthesis of global communication networks available at low cost, enormous growth in popular uptake of personal computers and communication devices, and the need for more sophisticated discussion of complex issues are continually pushing the boundaries of our expertise. Virtual learning communities (VLCs) are emerging constructs that depend on the notion of socially constructed learning to provide a focus for informed discussion and lifelong learning. They make use of increasingly sophisticated technologies to establish, support, and maintain communities—collections of individuals with a common purpose, acting in social settings, and geographically disparate.

Virtual learning communities are defined as groups of individuals that come together to study some area of common interest. They are virtual communities in the sense that they depend on a variety of information and communication technologies (ICTs) to coordinate their activities. They share many characteristics with virtual communities of practice. The nature of the relationships among these three constructs is explored in this article. The role of ICTs and multimedia in supporting VLCs is reviewed. This article concludes with a summary of the challenges facing both the organizations in stimulating the presence and growth of VLCs and the individuals who participate in such communities.

BACKGROUND

Virtual Communities

A virtual community is a cyber-location where a group of individuals can meet on the basis of a shared interest. Virtual communities are enabled by ICTs such as the Internet, the World-Wide Web, electronic mail, discussion forums, chat rooms, conference calls, and so on. Access may be restricted or unrestricted; activity or discussion may be moderated or unmoderated. Such communities may have a physical location, or they may be purely virtual.

Virtual Learning Communities

Learning communities form where individuals come together to study, often in connection with some formal course. Social constructivism is a process of learning where knowledge about a topic is actively constructed (Jonassen & Duffy, 1992) and where all participants in the community have a role to play in the development of knowledge (Jarvis, Holford & Griffin, 2003). This social aspect is central to the notion of a learning community where meaning is negotiated by the group as a whole. A learning community can capture the experience of current and prior participants and act as a resource for future ones.

A virtual learning community is a kind of virtual community where the motivation of group members is the study of some topic in order to learn or construct knowledge about it. Virtual learning communities extend traditional learning communities by meeting in spaces that have an online component. As with virtual communities of practice, virtual learning communities benefit from face-to-face contact (Kowch & Schwier, 1997).

When learning communities become virtual, the activities of inquiry and interaction are mediated by technology rather than face-to-face attendance. There are factors that affect social learning as a result; while the barrier of distance may be removed, the barriers of access and information literacy are raised instead. The virtual learning environment is a collection of tools and technologies that supports the activities of the community.

Communities of Practice

A community of practice is described as "a set of relations among persons, activity and world, over time and in relation with other tangential and overlap-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

ping CoPs" (Lave & Wenger, 1991, p. 98). Hildreth et al. (1998) described communities of practice that coordinate work in a geographically distributed sense.

A community of practice is characterized by "individuals with common expertise participating in an informal relationship to resolve a shared problem or situation that impacts upon their shared futures" (Bowles, 2002, reported in Kilpatrick, 2003). The construction of knowledge is enabled by a sophisticated process of negotiation and collaboration, and the social capital that results develops the understanding of professional practice. The characteristics of negotiation, collaboration, shared understanding, and shared interest are in common with virtual learning communities.

Virtual communities of practice are enabled by the same kinds of ICTs as virtual learning communities. Such constructs come into being to support collaboration among professionals across wide geographic distribution, for example. While they are enabled by ICTs, face-to-face interaction appears to be a crucial element of such communities in order to cement relationships and build trust among participants (Schwier, Campbell & Kenny, 2004).

MAIN FOCUS—LEARNING THROUGH INTERACTION

In a virtual learning community, problems, issues, and activities are defined by negotiation among participants (McConnell, 2004; Schwier, 2004). Participants build knowledge in a social setting and engage in discourse related to the purpose of learning in the chosen area. Virtual learning communities are thus socially-centered and task-oriented. Participants learn from each other by doing authentic tasks.

The patterns of interaction that can occur during the activities of a VLC are complex and vary with the nature of a task, the participants involved, and the technology used. Although VLCs depend on technology for their existence, any particular technology is viewed as a tool, not a central artifact. In this section, we review the role of different ICTs that can support VLCs and explore the types of social constructs that can emerge.

The Role of Technology in Mediating Interaction

Technology can provide support for different forms of leadership and styles of working found in group organizations. We consider groups to negotiate their position along several key dimensions (Table 1) that characterize the learning environment. Flexibility in the environment to support negotiation is a key factor in the effectiveness of the VLC.

The position occupied along each dimension affects the nature of activity in a virtual learning community. For example, if a group member undertakes activity in a private workspace, the others must trust their integrity and commitment to complete the task and report back to the group. Conflicts can arise when individuals adopt positions significantly different from those negotiated by the group. The group must record its rationale for taking positions and renegotiate them as needed.

Certain technologies can be deployed to support these positions. For example, the use of electronic mail, discussion forums, and shared workspaces can support an empowering leader (Hansson, 1999) to coordinate work by allowing the leader and delegates to accomplish work in their own timeframes. The agility of the group is determined partly by how they

Table 1. Dimensions associated with learning activity

Focus of activity	Group	↔	Individual
Working environment	Shared	↔	Private
Nature of interaction	Discourse	↔	Argument/rhetoric
Mode of interaction	Synchronous	↔	Asynchronous
Management method	Self-organized	↔	Delegated
Nature of leadership	Co-operative	↔	Traditional (power)
Style of Learning	Co-operative	↔	Collaborative

Table 2. Technologies for VLCs, affordances and affect on the learning process

Technology	Synchronous/ Asychronous	Affords	Affects/Affected by
Chat room	S	Multiple conversations	Participation due to ability required
Video (conference)	S	Enhanced social presence	Turn-taking, eye contact
Audio (conference)	S		Non-verbal communication
Video/animation (recorded)	A	Individual study; annotation; portrayal of action	Engagement with group
Hypertext/Web pages	A	Context; accessibility; delivery mechanism	Engagement
Shared collaborative space (whiteboard)	S	Multiple activity focused on objects; non-verbal interaction; less interference	Moderation
Electronic mail	A	Negotiation of schedules; task-oriented activity	Information load
Newsgroup / discussion forum	A	Multiple threaded discussions; reflection	
Portal	A	Access; sense of belonging and personal space	Technical constraints
Audio + shared space	S	Efficient combination of attention streams	Accessibility due to ability required
Simulation / visualization	Both	Explanation of complex phenomena	Technical constraints and quality

can assimilate into their environment a variety of technologies. Their ability to do this is often constrained by organizational factors (cost, availability), inflexibility (environment does not support), or capability (group member's ability to use the technology.)

The main challenge to supporting a learning community is in providing an environment to facilitate collaboration and communication. The challenge is made greater by the need to relegate technology into the background, allowing participants to concentrate on task-related activity. The role of technology in supporting activity is complex, and it is useful to consider each technology in terms of what potential it affords and how it affects the learning process or interaction (Table 2, synthesized from Barner-Rasmussen, 1999 and Dillenbourg, 2000).

Effective interaction is enabled by a blend of mutually supporting technologies that allow the user to concentrate on the task at hand. For example, audio, video, and text combined provide the greatest impression of social presence but are difficult to coordinate. For a task involving some object, a shared view of the object combined with an audio stream may be more effective than when combined with video. Video conferences have interaction issues due to the contention for visual attention, lack of eye-contact, and non-verbal cues for turn-taking. Animation, simulation, and visualization can help to explain complex phenomena but raise technical challenges with regard to equipment, communication, and the design and annotation of resources for learning.

Note that while hypertext and Web sites may be important techniques for delivery of information, they need to be enhanced to enable social collaboration and extension of knowledge created by others. The way in which resources are designed, created, and maintained within the environment raises interesting technical challenges as to how such annotations can be made and how they propagate through the community.

Several studies (Juhlin et al., 2001; Wenger, 1998; Heath & Luff, 1991) have shown that it is crucially important to be able to refer visually to what one is discussing. For example, Heath & Luff (1991) found that the decision-making capacity of London Underground managers was impaired when they could not see the shared situation board. This kind of interaction requires a shared workspace in a synchronous setting. For a group that engages in vigorous discourse, synchronous communication enabled via chat rooms or teleconferences are useful and can enhance the quality of learning (Mercer, 2003). However, for large groups, it can be hard for everyone to take part and keep track of the discussion.

Asynchronous technologies provide participants with time for reflection in their practice. Halverson & Ackerman (2003) provide an example of how an asynchronous mechanism can support a community

(in this case, a community of practice). They describe the evolution of an artifact (in this case, a document) that captures many facets of organizational memory constructed in a cooperative environment.

Face-to-face communication remains important, even with a virtual community. There are many reports (Hildreth et al, 1998; Isahaya & Macauley, 1999; Li & Williams, 1999) that interaction in the virtual world is enhanced after meeting and activity in real life (IRL). This is especially important for virtual communities, since building trust is a more complex process (Duarte & Snyder, 1999).

The Nature of Group Interaction

Trust is a key factor in the construction of social capital (Kilpatrick, 1999). This is the stuff that defines expected behaviors and values, fosters a sense of trust and shared values, and establishes communication paths. The nature of the groups and the characteristics of the individual members determine how such social capital is constructed. Constantine (1993) defines four classes of groups (Figure 1) and illustrates how their characteristics are suited to different classes of problems.

There are differences in the nature of activity, communication, and management style for each type of group. Some management styles and discussion formats are inappropriate for some groups (i.e., power-based leadership can affect the nature of open discussion). It seems that for the purposes of virtual learning communities, breakthrough and open collaboration teams are the most appropriate. In the former, creative activity takes place by individuals through their own inquiry in a subarea and later contributes to the work of the entire team. In the latter, the team functions as a whole, with each member actively contributing, discussing, and extending the work of others.

Subgroups can form or be established within virtual learning communities. Some of the following subgroups have wider relevance, but the discussion here is focused primarily on educational contexts.

- **The Cohort:** A group of peers come together because they share common characteristics, typically age and academic maturity, and because they engage in common activities of study. The cohort can be an extremely strong social force within a community.
- **Cliques and Factions:** A clique is a cohort that forms spontaneously but which excludes others that are not recognized by its members. The members of a faction are motivated by political or ideological influences and see themselves as diametrically opposed to other groups within the community. Both types of subgroups can have negative effects on the community through exclusion, alienation, and conflict.
- **Birds of a Feather:** Can form spontaneously where group members wish to pursue a particular area of specialization within the community.
- **The Loner:** Although a community can focus learning and activity, there will be individuals who either are not motivated to engage with the community or are unable to engage. The learning community must respect the rights of the individual to pursue private activities in association rather than by immersion.

The shared interest that virtual communities possess makes the first three of these groups likely to appear. The environment that supports the community needs to provide space for loners and mediate the negative effects of cliques and factions by negotiation.

Role of the Facilitator

Due to the nature of the primary communication mechanisms, leadership in virtual learning communities may be difficult to establish, and such communities have greater difficulty reaching decisions than those that meet face-to-face (Farnham et al, 2000). There is a role for a facilitator to guide and moderate discourse and to take on a wider leadership role in a virtual learning community.

The nature of a discussion may be affected by the personality mix of the contributing individuals. Task-oriented or self-oriented individuals may contribute information or insights directly related to the work at hand; interaction-oriented individuals may enjoy the actual activities of discussion and working with others. The role of the facilitator is to progress activity and discourse from shallow, trivial exchanges into deeper learning.

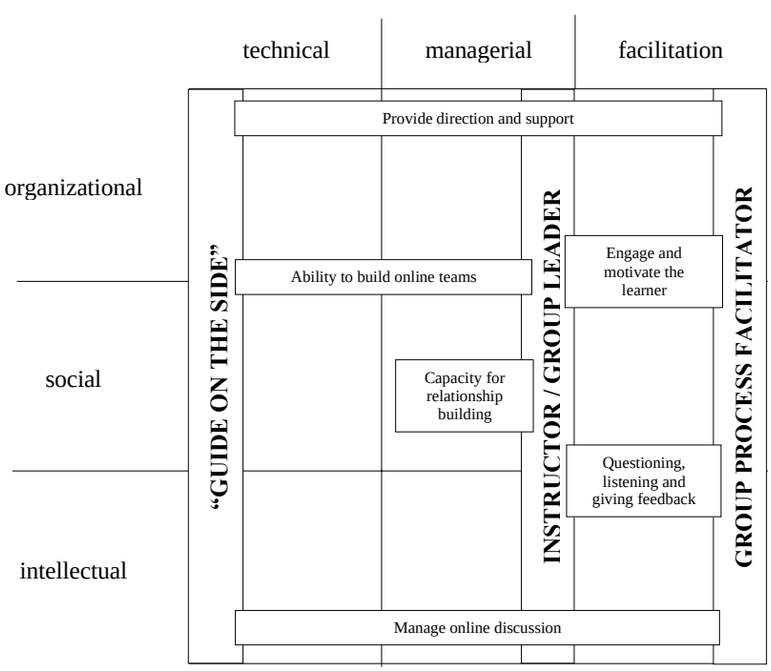
Figure 1. Types of groups and their characteristics

Random (breakthrough) bottom-up decision making promotes creativity chaotic, competitive operates in a chaotic, undirected way good for creative breakthroughs	**Open (problem-solving)** Decision making by consensus adapt to and solve complex problems chaotic, might not solve anything operates in a flexible, explorative way works best on solving complex problems
Synchronous (vision-sharing) decisions implied by visions efficient, smooth can drift, lose sight of reality operates in a coordinated way repetitive problems, high performance	**Closed (tactical)** traditional leadership stable and secure can be over-controlled operates in a group-oriented way good for routine projects

The facilitator acts as a friend and coach to the community, mediating the discussion, allowing issues to emerge, and coaxing where necessary. The facilitator has a complex technical role to play, often acting as a troubleshooter to resolve technical difficulties. The facilitator can benefit from tools that are constructed to support task-oriented activities (Bellotti et al, 2003; Takkinen, 2002) in carrying out its managerial role to schedule task-related activities and negotiate timetables among participants.

To show the various skills, roles, and activities associated with the facilitator, Figure 2 has been synthesized from Paulsen (1995), Kemshal-Bell (2001), Collision (2000), and Backroad Connections (2002). Skills are divided into organizational, social, and technical; activity is categorized into technical, managerial, and facilitation, and the roles of group process facilitator, instructor/group leader, and guide are shown to cut vertically across skill sets. Overlap between activities shown in boxes indicates that an activity is performed in a particular role.

Figure 2. Skills, functions, roles and activities of the facilitator



FUTURE TRENDS

Technical Challenges

It seems feasible to combine and use contemporary multimedia and ICTs to support virtual learning communities. Indeed, large-scale installation of learning platforms such as Blackboard (Blackboard Inc, 2002) and WebCT (WebCT Inc., 2003) indicate that many institutions have recognized the value of what such communities can offer. The current technical challenges are concerned with the design of course materials; the acquisition, deployment, and maintenance of suitable ICT infrastructure; and the design of learning environments to suit the task of collaborative learning. Table 3 lists some desirable characteristics for such environments.

An interesting challenge for multimedia development is in the area of annotation and rights management. To support the needs of social construction of learning, individuals must be able to synthesize their own materials from resources within and external to the environment. At a technical level, this may be hard to achieve or difficult for the individual to do; in terms of copyright and intellectual property, there are conflicts with controls imposed by digital rights management technologies.

Organizational Challenges

In virtual learning communities, activity is guided by the influence of participants on each other, not the nature of any power relationships that may exist. This can lead to a conflict within organizations that is based on hierarchies and authority relationships. Communities of practice face the same issues—many organizations cannot support them within their structure. Organizations, particularly established institutions of learning, have to face the challenge of deciding whether the value added by social and intellectual capital outweighs the difficulties in establishing and maintaining such communities.

Table 3. Desirable characteristics of learning environments

Desirable Characteristics of Learning Environments
• Readily accessible to all participants
• Promotes the principles of negotiation, intimacy, commitment, and engagement, and enables control by the participant
• Reflects the image of the organization to reinforce a sense of belonging
• Permits customization to reinforce sense of private and personal space
• Acts as a repository of organizational memory of current and past participants
• Archives, indexes, and enables searching of the repository to allow effective access to previously constructed knowledge
• Allows the repository to grow by incorporating annotations, discourse, and materials produced by participants
• Highlights and makes accessible the terminology (vocabulary) of the

The challenges for educational institutions are not only to adjust their structures in order to incorporate virtual learning communities, but to determine whether they can justify the level of expenditure necessary to provide the appropriate level of technology support. They have to resolve the problems associated with changes to the status and role of academic staff and the activities that they perform in working with virtual learning communities.

Educational Challenges

A virtual learning community requires strong leadership in order to become established. While this can be done initially by an academic leader, the desire is that new leaders will emerge as the community develops. The establishment and maintenance of a virtual learning community lead to changes in the role of teachers as facilitators of learning.

Modern trends in higher education point toward constructivist approaches that emphasize the importance of social learning. Learning communities facilitate construction and sharing of knowledge. In particular, Vygotsky (1978) proposed that the idea of a zone of proximal development—the level that an individual can attain in conjunction with a group of others—is most relevant as it indicates the potential for an individual to develop through interaction.

Virtual learning communities begin to meet the challenge of how to enable social learning in a mass education system while preserving the characteristics of task engagement and substantive discourse. The scale on which this is done affects the workload associated with creating and sustaining a virtual learning community. The changing role of staff and its workload are important factors in the growth of virtual learning communities.

CONCLUSION

Virtual learning communities are important, dynamic, and exciting constructs. They emerge to support a variety of activities related to learning and are sustained by their members as long as they are useful. Multimedia, information, and communication technologies play a strong supporting role in the establishment and maintenance of these communities.

The future development of such communities is driven by the increase in distance education and the trend of social learning in higher education. Institutions need to consider carefully the advantages and costs of establishing these kinds of communities to support learning. In wider terms, because of the overlap with virtual communities of practice, many of these challenges will be faced by organizations outside higher education.

REFERENCES

Backroad Connections Pty Ltd. (2002). Effective online facilitation. Australian National Training Authority.

Barner-Rasmussen, M. (1999). Virtual interactive learning environments for higher-education institutions. *Proceedings of the Nordic Workshop on Computer-Supported Collaborative Learning*, Göteberg, Sweden.

Belotti, V., Ducheneuat, N., Howard, M.A., & Smith, I.E. (2003). Taking e-mail to task: The design and evaluation of a task management centered e-mail tool. *Proceedings of ACM Conference on Human Factors in Computing Systems*, Fort Lauderdale, Florida.

Blackboard Inc. (2002). Blackboard learning system: Product overview white paper. Washington, D.C.

Bowles, M. (2002). *Forming a community of practice in North/NorthEast Tasmania on responsive and flexible VET*. Launceston, Australia: TAFE.

Collision, G., Erlbaum, B., Haavind, S., & Tinker, R. (Eds.). (2000). *Facilitating online learning: Effective strategies for moderators*. Madison: Atwood Publishing.

Constantine, L. (1993). Work practice and organization. *Communications of the ACM, 36*(10), 34-43.

Dillenbourg, P. (2000). Virtual learning environments. *Proceedings of the EUN Conference 2000 Workshop on Virtual Learning Environments*, Brussels, Belgium.

Duarte, D.L., & Snyder, N.T. (1999). *Managing virtual teams: Strategies, tools and techniques that succeed*. San Francisco: Jossey-Bass.

Farnham, S., Chesley, H.R., McGhee, D.E., & Kawal, R. (2000). Structured online interactions: Improving the decision-making of small discussion groups. *Pro-*

ceedings of the ACM Conference on Computer Supported Cooperative Work - CSCW 2000.

Hansson, H. (1999). Demands on virtual teams and virtual leadership to support sustainable learning processes. *Proceedings of the Nordic Workshop on Computer-Supported Collaborative Learning*, Göteberg, Sweden.

Heath, C., & Luff, P. (1991). Collaborative activity and technological design: Task coordination in London underground control rooms. *Proceedings of the Second European Conference on Computer-Supported Cooperative Work*.

Hildreth, P., Kimble, C., & Wright, P. (1998). Computer mediated communications and communities of practice. *Proceedings of the International Conference on the Social and Ethical Impacts of Information and Communication Technologies*, Erasmus University, The Netherlands.

Ishaya, T., & Macauley, L. (1999). The importance of social awareness in global information systems. *Proceedings of the 4th UKAIS Conference*.

Jarvis, P., Holford, J., & Griffin, C. (2003). *The theory and practice of learning*. London: Kogan Page.

Jonassen, D.H., & Duffy, T.M. (Eds.). (1992). *Constructivism and the technology of instruction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Juhlin, O., & Weilenmann. (2001). Decentralizing the control room: Mobile work and institutional order. *Proceedings of the 7th European Conference on Computer-Supported Cooperative Work*.

Kemshal-Bell, G. (2001). The online teacher. ITAM ESD, TAFENSW.

Kilpatrick, S., Barrett, M., & Jones, T. (2003). Defining learning communities. *Proceedings of the AARE International Education Research Conference*, Auckland, New Zealand.

Kilpatrick, S., Bell, R., & Falk, I. (1999). The role of group learning in building social capital. *Journal of Vocational Education and Training, 51*(1), 129-144.

Kowch, E., & Schwier, R.A. (1997). Characteristics of technology-based virtual learning environments. *Proceedings of the Second National Congress on Rural Education*, Saskatoon, Canada.

Lave, J., & Wenger, E. (1991). *Situated learning. Legitimate peripheral participation*. Cambridge, MA: Cambridge University Press.

Li, F., & Williams, H. (1999). Organizational innovations through information systems: Some lessons from geography. *Proceedings of the 4th UKAIS Conference*, York.

McConnell, D., Lally, V., & Banks, S. (2004). Theory and design of networked learning communities. *Proceedings of the Networked Learning Conference*, Lancaster, UK.

Mercer, D. (2003). Using synchronous communication for online social constructivist learning. *Proceedings of the Canadian Association for Distance Education Conference*, St Johns, Newfoundland.

Paulsen, M.F. (1995). Moderating electronic conferences. In Z.L. Berge, & M.P. Collins (Eds.), *Computer-mediated communication and the online classroom in distance learning*. Cresskill, NJ: Hampton Press.

Schwier, R.A. (2002). Shaping the metaphor of community in online learning environments. *Proceedings of the International Symposium on Educational Conferencing*, Banff, Canada.

Schwier, R.A., Campbell, K., & Kenny, R. (2004). Instructional designers' observations about identity, communities of practice and change agency. *Australasian Journal of Educational Technology, 20*(1), 69-100.

Takkinen, J. (2002). From information management to task management in electronic mail [unpublished Ph.D. dissertation]. Linköping, Sweden: Linköping University.

Vygotsky, L. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

WebCT Inc. (2003). WebCT product datasheet. Lynnfield, MA.

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. Cambridge: Cambridge University Press.

KEY TERMS

Community of Practice: A community of professional individuals who have the shared sense of purpose in a work situation (e.g., professionals at different institutions collaborating on best practice, or individuals that perform the same function in different parts of an organization).

Computer-Mediated Communication / Computer-Supported Collaborative Work: The use of information technology to support the interaction between people, directed to the resolution of a problem or activity in a task context.

Intellectual Capital: Knowledge or information that is created through collaborative activity by a community. It can be difficult to ensure clear notions of ownership since knowledge is jointly held. The organization of communities must define codes of behavior to deal with ownership issues.

Learning Community: A community that is established or which comes together with the purpose of studying or learning. Often created to support a course of study in real life, or can emerge spontaneously when common purpose or interest is identified.

Self-Organizing Group: A subset of a community where individuals decide for themselves how work structures, study activity, and so forth are coordinated. Such structures make use of social interactions and commonality to be sustained for as long as they are needed.

Social Learning: Process of constructing knowledge by individuals working in groups. A shared understanding emerges from individual understanding coupled with communication or collaborative exploration of an area of interest. This style of learning generates intellectual capital (knowledge that is jointly held) and social capital (trust, mutual respect).

Virtual Community: A social and technical construct that exists to coordinate the group-based activity of a number of individuals who share a common interest or sense of purpose. Virtual communities are maintained in the online world and supported by communication technology to support geographically separated groups.

Virtual Learning Community: Variant of the above where individuals come together, often in connection with a course of study or academic activity, to study or investigate problems related to a theme or area of shared interest.

Virtual Reality and HyperReality Technologies in Universities

Lalita Rajasingham
Victoria University of Wellington, New Zealand

John Tiffin
Victoria University of Wellington, New Zealand

INTRODUCTION

The term HyperReality (HR) was coined by Nobuyoshi Terashima to refer to "the technological capability to intermix virtual reality (VR) with physical reality (PR) and artificial intelligence (AI) with human intelligence (HI)" (Terashima, 2001, p. 4). HR is a technological capability like nanotechnology, human cloning and artificial intelligence. Like them, it does not as yet exist in the sense of being clearly demonstrable and publicly available. Like them, it is maturing in laboratories where the question 'if?' has been replaced by the question 'when?' And like them, the implications of its appearance as a basic infrastructure technology are profound and merit careful consideration (Tiffin & Terashima, 2001). Because of this, universities – if they are to be universities – will be involved with HR as a medium and subject of instruction and research, and for the storage and development of knowledge (Tiffin & Rajasingham, 2003). The concepts of HyperUniversities, HyperClasses, Hyperschools and HyperLectures are at the same level of development as the concepts of virtual universities, virtual classes, virtual colleges and virtual schools in the later part of the 1980s (Tiffin & Rajasingham, 1995).

HR subsumes virtual reality. HR is only possible because of the development of computer-generated virtual reality, in particular, the development of distributed VR, which makes it possible for different people in different places to interact together in the same VR. It was the theoretical application of this capability to education and especially to university education that led to the concept of virtual classes in virtual schools and universities (Tiffin & Rajasingham, 1995). Initial experiments simulated virtual classes by using videoconferencing, audio conferencing and audiographic conferencing. The emergence of the Internet shifted these ideas from a laboratory stage to development of institutions calling themselves virtual universities and virtual schools by virtue of being able to bring teachers and students together in classes using telecommunications and computers instead of public transport and buildings.

HR also subsumes AI. Teaching machines and computers have been used for instruction since the early days of CAI (Computer-Assisted Instruction) in the 1960s, albeit with little overall impact on education, especially at the university level. However, the growing capability and ubiquity of AI expert systems and agents, the vast amount of repetitive work involved in teaching and the growing application of business criteria to the management of education suggests that AI agents, conceivably in avatar form, will be adopted in education, and the place where this will begin is likely to be in universities.

THE NEED

Worldwide, governments face the challenge of increasing demand for university education. In Asia alone, the numbers seeking university places is predicted to rise from 17 million in 1995 to 87 million by 2020 (Rowe, 2003). It is unlikely that such demand can be fully met using the traditional communications systems of education (Daniel, 1996). These are:

- The public transport systems that bring students and teachers together for regular epi-

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

sodes of face-to-face instructional interaction called classes, lectures, seminars or tutorials.
- The buildings that provide dedicated instructional environments called classrooms, lecture theatres or seminar rooms characterised by frame-based presentation media and workspace on desks and tables. The buildings also need support environments such as offices, rest areas and recreational facilities.
- Provision for the use of paper-based storage media (books, notebooks, exercise books, assignment folders) in libraries, carrels, desks, assignment drops.
- Laboratory space and facilities.
- Infrastructures for telecommunications.

The costs of building and maintaining universities and the support infrastructures they need are high and getting higher. Increasingly, universities turn towards the Internet, where students and teachers can be brought together as telepresences in virtual classes, virtual lectures, virtual seminars and virtual tutorials. Rumble (1997, 1999, 2004), Turoff (1996) and Butcher and Roberts (2004) all agree that virtual universities on the Internet are significantly less costly than conventional building-based universities. Virtual universities that function primarily through the Internet and have no buildings for student needs and no demand on public transport infrastructures for students have been in existence since the mid-1990s. At minimum, conventional universities now have a home page on the Web; their students use the Web to help with assignments and to link with other students, teachers and administrators; and university management is exploring other ways of expanding teaching and administration activities on the Internet.

Initially, people tend to communicate through new media in the manner of the old media they are accustomed to. Universities use the Web as a library resource and for what was traditionally done by means of handouts and brochures, and e-mail for housekeeping notices, seminar discussion and written assignments. Virtual universities on the Internet tend to operate as electronic correspondence colleges. However, the Internet is becoming broadband, and computers get more powerful and portable. Universities can now use the Internet for streamed lectures and for holding classes by audiographic conferencing and video conferencing. It is possible for students and teachers to have telepresence as avatars and be fully immersed in three-dimensional distributed virtual classes (Tiffin & Rajasingham, 2001).

The Virtual Class/Lecture/Seminar

Roxanne Hiltz coined the term "virtual classroom" for the use of computer-generated communications "to create an electronic analogue of the communications forms that usually occur in a classroom, including discussion as well as lectures and tests" (Hiltz, 1986, p. 95). In 1986, John Tiffin and Lalita Rajasingham inaugurated a long-term action research program with postgraduate students at Victoria University of Wellington, New Zealand that sought to conduct what they called virtual classes, where students communicated with computers linked by telecommunications. They used the term "class" in the sense of an interactive instructional communication function between teachers and students and between students and the term "virtual" in the sense of existing in effect, but not in fact. Tiffin and Rajasingham hypothesized that learning could be effected by means of computers interlinked by telecommunications without the physical facts of classrooms, schools, colleges and universities. In contrast to Hiltz, they assumed that education delivered in this way would not be analogous to conventional educational practice, but would be modified by the new information technology and take new forms; and that in time this would include meeting for interaction in computer-generated virtual realities which would become increasingly immersive. They concluded that a virtual class need not necessarily be synchronous and that the people in it formed virtual networks that were independent of location. "The effect would be to make education available anywhere anytime" (Tiffin & Rajasingham, 1995, p. 143).

The "virtual class" research project began in pre-Internet days of 1986 using a lash-up of equipment that sought to comprehensively conceptualise what would be involved in a virtual university that depended on computers and telecommunications. Assignments, student-to-student and student-to-teacher discourse and course administration were online, and a variety of audiographic modes were developed

for lectures, seminars, tutorials, tests and examinations.

The project linked students and teachers at national and international levels, and it became apparent that the convergence and integration of computers and telecommunications in universities and schools had a dynamic of its own. Experimentation with audioconferencing, audiographic conferencing and video conferencing systems was taking place worldwide, usually as an initiative of individual teachers (Rajasingham, 1988; Acker, Bakhshi & Wang, 1991; Underwood, 1989; Donald, 1989; http://calico.org/chapters). Such activities multiplied with the coming of the Internet.

In 1995, Tiffin and Rajasingham published *In Search of the Virtual Class: Education in an Information Society*, which outlined the way virtual classes could serve as the basis for virtual schools, virtual colleges and virtual universities; and these now began to appear on the Internet. Inevitably the meaning of the terms began to change to reflect actual practice. The terms virtual schools, colleges and universities are now synonymous with the more recent terms, e-learning, e-schools, e-colleges and e-universities, which came into vogue with the commercialization of the Internet and the introduction of the concept of e-commerce. Essentially, these terms now refer to schools, colleges and universities that exist on the Internet.

Virtual universities are now beginning to appear in developing countries. The African Virtual University began operating in 1997 (www.col.org) and now has 31 learning centers at partner universities in 17 African countries. In 2003, 23,000 Africans were enrolled in courses such as journalism, languages and accounting. The Commonwealth of Learning is currently developing virtual universities for the Small States of the Commonwealth (www.col.org), and in 2003, the United Nations launched the Global Virtual University of the United Nations University (www.globetechnology.com/servlet/story/RTGAM.20030618.wun0617/BNStory/Technology). Being virtual on the Internet makes it possible for a university to market globally (Tiffin & Rajasingham, 2003). From the perspective of the World Trade Organization, universities provide an information service that should be freely traded as part of a process of globalization.

HR and the HyperClass (HC)

Even if it is more economic, the development of a virtual dimension to universities does not imply that they will cease to exist in physical reality. Many universities that have sought to exist solely on the Internet have found that students want some part of their education in PR. What we could be seeing is the development of a global/local hybrid university that exists in virtual and physical reality on the Internet and in buildings serving global needs and local needs. A technology that allows this duality is HR.

Developed in Japan's Advanced Telecommunications Research Laboratories under the leadership of Nobuyoshi Terashima, HR is a platform being developed for broadband Internet. HR permits the seamless interaction of VRs with PRs and HI with AI (Terashima, 2001). Jaron Lanier has since developed a similar concept of intermeshing PRs and VRs, which he calls Tele-immersion (Lanier, 2001). However, this does not allow for the interaction of AI and HI.

Working with Terashima from 1993 on the application of HR to education, Tiffin and Rajasingham coined the concept schemata HyperClass, HyperSchool, HyperCollege and HyperUniversity (2001) to describe an educational environment in which physically real students, teachers and subject matter could seamlessly interact with virtual students, teachers and subject matter, and AI and HI could interact in the teaching/learning process. What makes this possible is a coaction field, which "provides a common site for objects and inhabitants from PR and VR and serves as a workplace or activity area within which they interact" (Terashima, 2001, p. 9). Coaction takes place in the context of a specific domain of integrated knowledge. So a coaction field could be a game played between real and virtual people, or a real salesperson selling a car to virtual customers (who and what is real and who and what is virtual depends on the kind of perspective of self that exists in a telephone conversation). A HyperClass is a coaction field in which physically real students and teachers in a real classroom can synchronously interact in a joint learning activity that involves a clearly defined subject domain with virtual students and teachers in other classrooms in other

universities in other countries. The first experimental HyperClass took place in 2000 between teachers and students at Waseda University and Victoria University in Japan. To the people in Japan, the New Zealanders were virtual; to the people in New Zealand, the Japanese were virtual. The subject was antique Japanese ceramics, and virtual copies of these were passed back and forth between the two classrooms that made up the HyperClass (Terashima, 2001).

A HyperClass creates a common space to reconcile learning that is local with learning that is global. It can be conducted in more than one language and holds the possibility of understanding a subject from the multiple perspectives of different cultures using text, aural and three-dimensional visual modes of communications (Tiffin & Rajasingham, 2001, 2003).

JITAIT

The HyperClass enables communication and interaction between PR and VR, but what could have even more impact on universities is that it provides a platform for communication between HI and AI. Applying HR to education means applying AI to education and designing a pedagogical interaction between HI and AI.

At the heart of the Vygotskyian approach expressed in the Zone of Proximal Development (1978) is the idea that when learners have difficulty in applying knowledge to a problem, they will learn more effectively if they can turn to someone in the role of teacher who can help them. This is the fundamental purpose of education; yet in the modern school and university, teachers are only available to respond to student needs during fixed hours, and even then, they have to share their attention with large groups of students. In principle, an artificially intelligent agent can be available whenever they are needed. Hence, the idea of just-in-time artificial intelligent tutors (JITAITs). In a university, they would be expert in a particular subject domain, endlessly learning from frequently asked questions, and available anytime and anywhere to deal with the more repetitive functions of teaching (Tiffin & Rajasingham, 2003).

CONCLUSION

There is growing disjuncture between the demand for university education and the capacity of conventional universities to respond.

The modern university is based on building and transport technologies and becomes increasingly costly. There has to be a way that is more economical and efficient, more matched to the times and technologies we live with, more open to people with languages other than English and more concerned with the curricula needs and cultural concerns of globalization that is available to anyone throughout their lives. Virtual universities have appeared in response to this, and conventional universities are developing virtual global dimensions on the Internet. But the Internet is becoming broadband, and computers are becoming more powerful and portable. Universities could become a hybrid mixture of the traditional place-based institutions that we know and that address local needs, and as cyberbased businesses that address global markets. An emergent technology that addresses this is HR, which could see HyperClasses in HyperUniversities that incorporate the use of JITAITs.

REFERENCES

Acker, S., Bakhshi, S., & Wang, X. (1991). User assessments of stereophonic, high bandwidth audioconferencing. *ITCA Teleconferencing Yearbook* (pp. 189-196).

Butcher, N., & Roberts, N. (2004). Costs, effectiveness, efficiency: A guide for sound investment. In H. Perraton, & H. Lentell (Eds.), *Policy for open and distance learning*. London: RoutledgeFalmer.

Daniel, J. (1996). *Mega-universities and knowledge media: Technology strategies for gigher education.* London: Kogan Page.

Donald, C. (1989, December). Technology convergence under Windows: An introduction to object oriented programming. Retrieved May 15, 2004, from *http://calico.org*

Hiltz, R. (1986). The virtual classroom: Using computer mediated communications for university teaching. *Journal of Communications, 36*(2), 95.

Lanier, J. (2001). Virtually there. *Scientific American*, April.

Rajasingham, L. (1988). *Distance education and new communications technologies.* New Zealand: Telecom Corp.

Rowe, M. (2003). Ideal way to lighten the load. Retrieved April 22, 2004, from *www.thes.co.uk/archive/story.asp?id=91859&state_value=Archive*

Rumble, G. (1997). *The costs and economics of open and distance learning.* London: Kogan Page.

Rumble, G. (Ed.). (2004). *Papers and debates on the economics and costs of distance and online learning.* Germany: Bibliotheks-und Informationssystem der Universität Oldenburg.

Terashima, N. (2001). The definition of HyperReality. In J. Tiffin & N. Terashima (Eds.), *HyperReality: Paradigm for the third millennium* (pp. 4-24). London; New York: Routledge.

Tiffin, J., & Rajasingham, L. (1995). *In search of the virtual class: Education in an information society.* London; New York; Canada: Routledge.

Tiffin, J., & Rajasingham, L. (2001). The HyperClass. In J. Tiffin & N. Terashima (Eds.), *HyperReality: Paradigm for the third millennium* (pp. 110-125). London; New York: Routledge.

Tiffin, J., & Rajasingham, L. (2003). *The global virtual university.* London; New York; Canada: Routledge.

Turoff, M. (1996). Costs for the development of a virtual university. Invited paper for Web-based Teleteaching '96, which is a component of the IFIP's annual meeting in Australia.

Underwood, J. (1989). Hypermedia: Where we are and where we aren't. Retrieved May 19, 2004, from *http://calico.org*

Vygostky, L.S. (1978). *Mind in society: The development of the higher psychological processes.* Cambridge, MA: Harvard University Press.

(1999). The costs of networked learning: what have we learnt? Paper to Sheffield Hallam University FLISH Conference. Retrieved March 23, 2000, from *www..shu.ac.uk/flish/rumblep.htm*

KEY TERMS

HyperClass, HyperLecture, HyperSeminar, HyperTutorial: Classes, lectures, seminars and tutorials that take place in a coaction field in HyperReality. This means an interaction between virtual teachers and students and objects and physically real teachers and students and objects to learn how to apply a specific domain of knowledge. It allows for the use of artificially intelligent tutors. Such systems are currently experimental, but have the potential to be used on the Internet.

HyperSchool, HyperCollege, HyperUniversity: The term Hyper means that these institutions could exist in HyperReality. HyperReality is where virtual reality and physical reality seamlessly intersect to allow interaction between their components and where human and artificial intelligences can communicate. The technological capability for this is at an experimental stage, but could be made available with broadband Internet.

JITAITS: Just-In-Time Artificially Intelligent Tutors are expert systems available on demand in HyperReality environments to respond to frequently asked student questions about specific domains of knowledge.

Virtual Class, Virtual Lecture, Virtual Seminar, Virtual Tutorial: Classes, lectures, seminars and tutorials are communication systems that allow people in the relative roles of teachers and learners to interact in pursuit of an instructional objective and to access supporting materials such as books and blackboards. The use of linked computers makes it possible for such interaction to take place without the physical presence of teachers and learners or any instructional materials or devices such as books and blackboards. The Internet now provides a global infrastructure for this so that the terms have become synonymous with holding classes, lectures, seminars and tutorials on the Internet.

Virtual School, Virtual College, Virtual University: The term virtual refers to the communication capabilities of these institutions and implies that they can be achieved by means of computers linked by telecommunications which, in effect today means by the Internet. The term "virtual" is used to contrast the way communications in conventional schools, colleges and universities requires the physical presence of teachers and learners and instructional materials and invokes the use of transport systems and buildings.

Web Content Adaptation Frameworks and Techniques

Tiong-Thye Goh
Victoria University of Wellington, New Zealand

Kinshuk
Massey University, New Zealand

INTRODUCTION

Most Web pages are designed with desktop platform access in mind, but with the proliferation of mobile devices such as Personal Digital Assistants (PDAs) and mobile phones, accessing Web pages through a variety of devices without proper content adaptation can result in an aesthetically unpleasant, un-navigable and, in most cases, unsatisfying experience. This article provides an overview of approaches in Web content adaptation framework and techniques being developed to extend the Web application access to non-desktop platforms. After describing general adaptation techniques, the article focuses particularly on the adaptation requirements of learning systems, especially when they are accessed through mobile devices.

WEB CONTENT ADAPTATION

Since most existing Web applications are geared towards desktop platforms, they limit the access only to a certain class of users, hence restricting the potential customer growth of the enterprise. With the increasing proliferation of a diverse set of devices accessing the Web under different network conditions, the need for content adaptation is significantly increasing. To circumvent this problem, various commercial products and research prototypes dealing with Web content adaptations have emerged, such as Spyglass (Spyglass, 2001), Intel QuickWeb (Intel, 1998), IBM Transcoding proxy (Smith et. al. 1999), Digestor (Bickmore & Shilit, 1997), Mobiware (Angin, Campbell, Kounavis & Liao, 1998), TranSend (Fox et. al 1998a), WingMan (Fox et.al. 1998b) and Power Browser (Buyukkokten, Garcia-Molina, Paepcke & Winograd, 2000). The types of content adaptation these systems looked into are mostly multimedia rich transformation. In contrast, there are other areas, such as mobile learning, which require the development of Web content adaptations for mobility with respect to user environment and capabilities. These areas have distinct features not yet researched extensively. This article provides an overview of some of the promising frameworks and techniques in content adaptation.

RE-AUTHORING

According to Bickmore and Schilit (1997), one straightforward method for content adaptation is to re-author the original Web content. Manual re-authoring can be done, but obviously, it is the most ineffective way and requires that the Web pages be accessible for re-authoring. This sometimes poses some practical constraints. However, the underlying principles and questions faced are identical for both automatic and manual re-authoring: What are the strategies used to re-author the pages? What are the strategies used to re-designate the navigations? What presentation styles can be achieved? These questions face any content adaptation process. The underlying principle is to isolate and distinguish the Web content objects, presentation objects, navigation objects and interactive objects for desktop publication and re-map them into other device-capable objects. Figure 1 shows such a re-mapping process. Once the strategies have been defined and the process matured, manual re-authoring can be converted into automated re-authoring through HyperText Transfer Protocol (HTTP) proxy server or server side techniques, such as common gateway interface (CGI), Servlet or client

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

side scripting. The re-authoring approach can either be mobile-device specific or tailored to multiple classes of devices. For multiple devices re-authoring, transformation style sheets (XSLT) and cascading style sheets (CCS) can also be used.

From another perspective, re-authoring can be viewed along two dimensions: syntactic (structure) vs. semantic (content), and transformation (convert) vs. elision (remove). Syntactic techniques operate on the structure of the page, while semantic techniques rely on the understanding of the content. Elision techniques basically remove some information, leaving everything else untouched, while transformation techniques involve modifying some aspect of the page's presentation or content. The Digestor system (Bickmore and Schilit, 1997) used the re-authoring technique that included outlining, first sentence elision and image scaling, and built an abstract syntax tree to provide content adaptation. The Digestor system used a proxy-based heuristic approach for its automated re-authoring. This method worked well for small-screen mobile devices. However, the elision process might remove certain content and affect the capturing of a user profile. There is also a possibility of making customization less accurate.

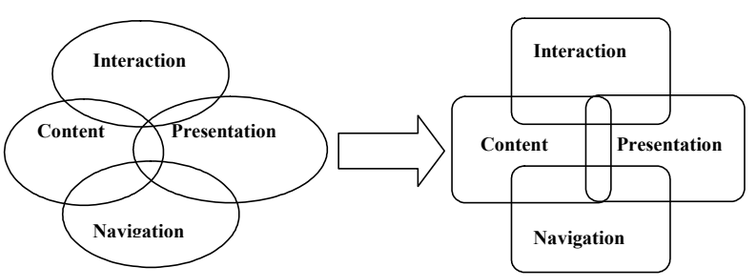
TRANSCODING

According to Bharadvaj, Anupam and Auephanwiriyakul (1998), modifying the HTTP streams and changing its content in situ is called "active transcoding" and is done dynamically without user intervention. Transcoding can be performed in both upstream and downstream directions. An implementation of this technique is MOWSER (Mobile Browser Project, 1996). MOWSER is an Apache proxy server agent written in Perl. MOWSER used proxy to perform transcoding. The incoming HTTP stream is modified by the proxy to include the capabilities and preferences of mobile users. The users' preferences and capabilities are stored in the server. Modification and update of preferences is done by a CGI form on a URL at a Web site maintained by the proxy. The proxy then fetches the files with the most suitable format to the requesting client. This implementation assumed that different formats are available for content adaptation. This is not an issue, as different formats can be created on the fly and cached in the server for future requests. Transcoding of images and videos is done using scaling, sub-sampling or sub key-frame techniques. Transcoding of HTML pages are done by eliminating unsupported tags and allowing users to select their preferences. This implementation, however, did not touch on the aspect of navigation. This technique, therefore, might not work well if adaptive navigation is required.

ANNOTATION-BASED CONTENT TRANSCODING

Annotation is a way to provide hints that enable a transcoding engine to make better decisions on con-

Figure 1. Desktop Web objects re-authored into mobile device-capable objects



tent adaptation (Hori, Kondoh, Ono, Hirose, & Singhal, 2000). This method uses some predefined "descriptor or syntax" to define the "rules" for transcoding or adaptation. An external file for the "descriptors" is recommended, as it separates content from the HTML markup tags. Annotation plays the role of a mediating representation, which provides semantics to be shared between meta-content authors and a content adaptation engine. A potential advantage of an annotation-based transcoding approach is the possibility of content adaptation based on semantics that cannot be achieved by approaches based on Web document syntax. Again, the fundamental principles discussed in re-authoring also apply in annotation-based transcoding, which comprises decomposition (isolation), combination (re-mapping) and partial replacement of content (distillation or elision). Hori et al. (2000) used the resource description framework (RDF) for implementing the annotation descriptors, and Xpath and Xpointer for associating an external description with a portion of an existing document. In their implementation, the relation between the HTML document and the annotation files was not limited to a one-to-one relationship. The annotation used predefined vocabulary, such as alternative, splitting hints and selection criteria. The problem with annotation-based transcoding is that it is very task oriented, and customization using a markup language is limited. It is also difficult to generalize. This method, however, is consistent in using Extensive Markup Language (XML)-related technologies.

MEDIA-RELATED RESOURCE FRAMEWORK

This approach defines a Web content adaptation framework using a definition identified as "related resource" (Lemlouma & Layaida, 2001). Related resources define a set of binary relations that can exist between a pair of media resources. A media resource can be an image, text file, audio file, video stream or something similar, and can be used by more than one document or application. A resource can be authored locally or imported, and may be used by the local server or a remote server. It can also be obtained after applying some transformation techniques. A relation exists between two resources and helps the adaptation process.

The architecture of the framework is comprised of the following entities:

1. **The server of content:** It maintains the multimedia content. Services may contain many heterogeneous media, such as text, video stream and audio, and may be authored in different versions. A server can use the content of another server belonging to the same multimedia system. Transformation is needed if the document is not in XML. A resource profile in Composite Capability/Preference Profile (CC/PP) structure is also needed (Butler, n.d.).
2. **Clients:** Clients have several characteristics, and they request their service demands to the servers of content. The clients' profiles are coded in XML/RDF as per CC/PP.
3. **A connection network:** Connection network ensures continued communication between the servers and the clients. No assumption is posed on the connection bandwidth, latency and accuracy. This means that the client and servers may interact in bad conditions, which must be taken into account when delivering multimedia content.
4. **Intermediate proxies:** Proxies can exist between the clients and the servers. A proxy may play the role of a client when the considered interaction is proxy-server oriented, and the role of a server when the considered interaction is proxy-client oriented.

The physical architecture is similar to the content adaptation structures in Bharadvaj et al. (1998) and Chen, Yang and Zhang (2000), while the conceptual model is similar to the object-oriented approach.

The framework provides considerations on Web content adaptation. The use of CC/PP and RDF with an XML-based structure is also consistent with recommendations and standards of the World Wide Web Consortium (W3C).

ADAPTIVE WEB CONTENT DELIVERY (AWCD)

The AWCD (Chen et. al., 2000; Ma, Bedner, Chang, Kuchinsky & Zhang, 2000) is perhaps the most

complete and practical approach to the adaptive Web content delivery. The goal is to improve content accessibility and perceived quality of service for information access under changing network and viewer conditions. This approach differs from other frameworks in the following ways:

1. It provides a decision engine for determining when and how to adapt the content.
2. It provides an automatic measure of network bandwidth availability and load for quality of service adaptation.
3. User preferences are registered through a Web form.
4. It is able to track session and user browsing behavior.
5. It is modular and extensible.

This approach lacks the following features that are desirable for any desktop or non-desktop platform adaptation:

1. CC/PP features are absent in this implementation.
2. Focus is on image and video adaptation, and the framework lacks "interactivities" and "navigation" or "intention" adaptation.
3. XSL (XSLT) is not used in the adaptation.
4. Offline adaptation is not considered in the framework, which is a serious limitation for non-reliable network bandwidth conditions.

FUNCTIONAL-BASED OBJECT MODEL (FOM)

FOM attempts to understand authors' intentions by identifying an object function instead of semantic understanding (Chen, Zhou, Shi, Zhang & Wu, 2001). It takes an overview of the Web site instead of being based on purely semantic structure information from HTML tags. The rationale is that every object in a Web site serves a particular function (basic and specific function) that reflect authors' intentions towards the purpose of the object. Based on this concept, a Web site is structured into an FOM model and adaptation rules are applied on the model. FOM includes two complementary parts: Basic FOM, based on the basic functional properties of objects; and Specific FOM, based on the category of objects.

A basic object is the smallest element in a hypermedia. It has both functions and properties and can be represented as follows:

- BO (Presentation, Semanteme, Decoration, Hyperlink, Interaction)
- Basic objects can be grouped into a composite object. A composite object has functions and properties and can be represented as follows:
- CO = {Oi, Clustering Relationship, Presentation Relationship | Oi is the Root Children of the CO, i=1, 2, ..., NR} where NR is the total number of Root Children of the CO.

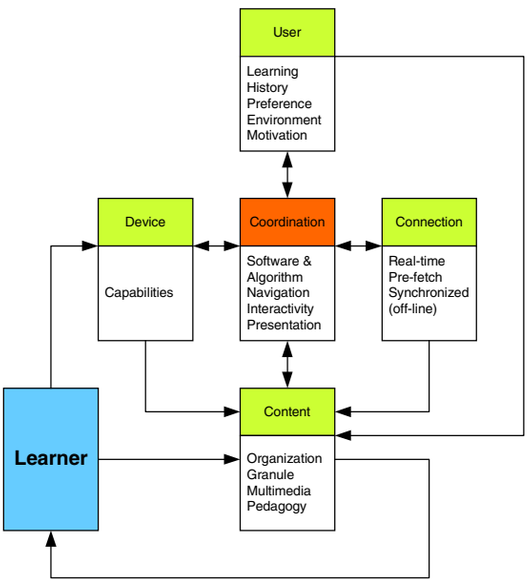
The specific function of an object in a given application environment is represented by its category, which directly reflects the authors' intentions. There are many object categories according to various purposes, such as information object, navigation object, interaction object, decoration object, special function object and page object.

This method requires basic object detection to be performed first to generate the necessary basic objects and category objects. Composite objects are detected by layout analysis of the Web pages using image pattern detection algorithms. Content adaptation rules are applied to these FOM, and the adapted pages are produced. There are separate rules for each object type.

MOBILE ADAPTATION FRAMEWORK

The mobile adaptation framework (Goh & Kinshuk, 2002; Kinshuk & Goh, 2003; Goh et al., 2003; Kainulainen et al., 2003) adapts content for users from mobility- and learning-centered perspectives. One of the unique characteristics of a mobile learner from a pedagogical perspective is the urgency towards the content delivery when and where the learner needs. Once the urgency has been detected, the system packages content suitable for such a condition. Another characteristic of a mobile learner is the mobility of learning settings. As mobility increases, the learning environment can be any-

Figure 2. Inter-relationship within the mobile adaptation framework



where, such as a hot spot, Internet café, classroom, campground, train or bus.

Adaptive systems are needed to facilitate the various requirements resulting from the array of mobile clients currently available, without duplicating the services and content on the server side. In case of mobile learning system adaptation, several dimensions of adaptation need to be considered. The mobile adaptation framework consists of five core competency dimensions (Goh & Kinshuk, 2002): content dimension, user model dimension, device dimension, connectivity dimension and coordination dimension. Within these dimensions exist sub-dimensions. Figure 2 depicts the interrelationships.

Content Dimension

This dimension represents the actual context and knowledge of the application. The *course modules organization* sub-dimension includes attributes such as part, chapters and sections of the content. Another sub-dimension is the *granular level* of the content, which indicates the level of difficulties of the content presented to the learner. *Multimedia* sub-dimension within the content dimension represents the multimedia representation of the content. This includes the use of text, audio, animation, video, 3-D video, animation and so on to represent the content to the learner. The *pedagogy* sub-dimension represents the teaching models and domain expert models that the system adopts.

User Dimension

The user dimension includes the attributes of the users. The *learning model* sub-dimension includes attributes such as module completed, weight and score, time taken, date of last access and so on, depending on the algorithms used in determining the learner profile. The *user preference* sub-dimension contains attributes such as preferred difficulty level and learning style. The *environmental* sub-dimension represents the actual location where the learner uses the system. Different environments, such as an Internet café, hot spot and classroom situation will have to be adopted differently. The adaptation must take into account the *motivation* sub-dimension, such as urgency of use.

Device Dimension

The device dimension consists of the *capabilities* sub-dimension, which includes attributes such as media support types and its capabilities in presenting multimedia content, display capability, audio and video capability, multi-language capability, memory, bandwidth, cookies, operation platform and so on. Adaptation depends on the sub-dimension in which the device is used.

Connectivity Dimension

Under this dimension, there are four operating sub-dimensions. The user can operate in a *real-time online* sub-dimension mode. In this aspect, the operating connecting speed and throughput determine some of the adaptation capabilities, such as a multimedia representation or text-based representation. Another sub-dimension is the *pre-fetching capability* of the application. While static pages can be pre-fetched easily, interactive applications need further consideration, such as the depth of pre-fetching. Here, device capability, network reliability and con-

necting type are the main considerations for adaptation. The third sub-dimension is *off-line synchronization*. Here, the attributes of depth and encrypted cookies need to be considered to provide seamless adaptation, especially for Web-based learning applications, which are highly interactive and where parameters regarding users' actions need to be returned to the server.

Coordination Dimension

The coordination dimension represents the *software and algorithm* sub-dimension of the application, the *presentation* sub-dimension, the *interactivity* sub-dimension of the application and the *navigation* sub-dimension. In any adaptive system, these dimensions must be well coordinated to provide users a good learning experience. The *software and algorithm* sub-dimension contains the script language and server page language to control the flow of the application from feedback through interactivity and navigation sub-dimensions. The *presentation* sub-dimension links the display and transformation of the content to the user. The *interactivity* sub-dimension represents how the user information can be sent back as feedback to the application. The *navigation* sub-dimension provides both feedback and movement within the application. For instance, in the connectivity dimension mentioned earlier, when a user operates under a synchronization sub-dimension, certain interactivities in the coordination dimension have to be dropped, and cookies and a script must be activated to store interactivity information, such as answers to a test. The coordination dimension provides these adaptations.

CONCLUSION

This article discussed several techniques and frameworks for Web content adaptation, ranging from the basic re-authoring to the more sophisticated frameworks that try to induce Web page object "intention." All these methods use similar underlying principles of trying to isolate content, presentation, navigation, interaction and intention. Once these components or objects have been identified, adaptation rules are applied to provide adapted content. Some of these methods provide network bandwidth measurement (Cheng et al., 2000; Bharadvaj et al., 1998; Goh & Kinshuk, 2002) to enhance quality of service or content adaptation. This parameter is important in the mobile environment, as the bandwidth is normally limited and connection easily interrupted. In the worst-case scenario, a pre-fetch or pre-sync method should automatically be recommended in place of online access. The situation is somewhat different when it comes to learning systems, particularly the access to learning systems through mobile devices, because none of these methods use domain knowledge, such as the specific features of the Web sites, to provide adaptation. Thus, by analyzing the key features of typical learning systems such as "interactivities" components and "navigation" components, we can prioritize these components and provide better adaptation that suits mobile environments. Pre-fetch or off-line access represent another mode of changing environment (bandwidth), which has not been covered in most of the frameworks, except Goh and Kinshuk (2002).

REFERENCES

Angin, O., Campbell, A. T., Kounavis, M.E., & Liao, R.F. (1998, August). The Mobiware Toolkit: Programmable support for adaptive mobile networking. *IEEE Personal Communications,* (4), 32-43.

Bharadvaj, H., Anupam, J., & Auephanwiriyakul, S. (1998, October). An active transcoding proxy to support mobile Web access. *The 17th IEEE Symposium on Reliable Distributed Systems,* West Lafayette, Indiana, October 20-23.

Bickmore, T.W., & Shilit, B.N. (1997), Digestor: device-independent access to the World-Wide-Web. *Proceedings of the Sixth International WWW Conference,* Santa Clara, California, April 7, (pp. 655-663).

Butler, M.H. (2002). Using capability classes to classify and match CC/PP and UAProf profiles. Retrieved October 18, 2002, from *www-uk.hpl.hp.com/people/marbut/capClass.htm*

Buyukkokten, O., Garcia-Molina, H., Paepcke, A., & Winograd, T. (2000). Power bBrowser: Efficient Web browsing for PDAs. *Proceedings of CHI2000* (*The Hague*, April).

Chen, J., Yang, Y., & Zhang, H. (2000, August). An adaptive Web content delivery system. *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-based Systems (AH2000)*, Trento, Italy, August 28-30.

Chen, J., Zhou, B., Shi, J., Zhang, H., & Wu, Q. (2001, May). Functional-based object model towards Web site adaptation. *Proceedings of WWW10*, (pp. 1-5). Retrieved October 18, 2002, from *www10.org/cdrom/papers/296/*

Fox, A., Goldberg, I., Gribble, S.D., Lee, D.C., Polito, A., & Brewer, E.A. (1998b, September). Experience with top gun wingman, a proxy-based graphical Web browser for the USR PalmPilot. *Proceedings of the IFIP International Conference on Distributed Systems Platforms and Open Distributed Processing (Middleware '98),* Lake District, UK.

Fox, A., Gribble, S.D., Chawathe, Y., & Brewer, E.A. (1998a). Adapting to network and client variation using active proxies: lessons and perspectives. *IEEE Personal Communication,* (4), August, 10-19.

Goh, T., & Kinshuk, T. (2002). A discussion on mobile agent-based mobile Web-based ITS. In Kinshuk, R. Lewis, K. Akahori, R. Kemp, T. Okamoto, L. Henderson & C.-H. Lee (Eds.), *proceedings of the International Conference on Computers in Education, Los Alamitos, CA: IEEE Computer Society* (pp. 1514-1515).

Goh, T., Kinshuk, T. & Lin, T. (2003). Developing an adaptive mobile learning system. In K.T. Lee & K. Mitchell (Eds.), *Proceedings of the International Conference on Computers in Education 2003,* Hong Kong, December 2-5 (pp. 1062-1065). Norfolk, VA: AACE.

Hori, M., Kondoh, G., Ono, K., Hirose, S., & Singhal, S. (2000, May 15-19). Annotation-based Web content transcoding. *Proceedings of the Ninth International World Wide Web Conference*, Amsterdam, Netherlands.

Intel QuickWeb. (1998). Retrieved from *www.intel.com/pressroom/archive/releases/IN011998.HTM*

Kainulainen, V., Suhonen, J., Sutinen, E., Goh, T., & Kinshuk (2003). Mobile digital portfolio extension. In J. Roschelle, T.W. Chan, Kinshuk, & S.J.H. Yang (Eds.), *Proceedings of the Second IEEE International Workshop on Wireless Mobile Technology in Education,* JungLi, Taiwan, March 23-25 (pp. 98-102). Los Alamitos, CA: IEEE Computer Society Press.

Kinshuk, & Goh, T.T. (2003, September 17-19). Mobile adaptation with multiple representation approach as educational pedagogy. *Sixth Business Informatics International Congress*, Dresden, Germany.

Lemlouma, T., & Layaïda, N. (2001, August). *A framework for media resource manipulation in an adaptation and negotiation architecture.* OPERA Project, INRIA Rhône Alpes.

Ma, W-Y., Bedner, I., Chang, G., Kuchinsky, A., & Zhang, H.J. (2000). A framework for adaptive content delivery in heterogeneous network environments. HP Lab. Retrieved October 18, 2003, from *www.cooltown.hp.com/dev/wpapers/adcon/MMCN2000.asp*

Mowser Project (1996). Mowser – A Web browser for mobile platforms. Retrieved October 18, 2003, from *www.cs.purdue.edu/research/cse/mobile/mowser.html*

Smith, J.R., Mohan, R. & Li, C. (1999). Scalable multimedia delivery for pervasive computing. *Proceedings ACM Multimedia 1999,* Orlando, October, (pp. 131-140).

Spyglass-Prism. (2001). Retrieved October 18, 2003, from *www.opentv.com/support/primer/prism.htm*

KEY TERMS

Annotation: A technique for content adaptation. Special tags are added to the HTML page to allow browsers to operate in a pre-defined function.

Apache: An open-source HTTP server for operating systems, including UNIX and Windows NT. A project supported by the Apache Software Foundation.

CGI: Common Gateway Interface is a standard protocol for users to interact with applications on Web servers.

HTML: HyperText Markup Language. The language used to create Web content.

HTTP: HyperText Transfer Protocol is a protocol for transferring requests and files over the Internet.

Proxy: A server that sits between the Web server and the client to provide protection and filtering.

RDF: Resource Description Framework. RDF is a language for representing information about resources in the World Wide Web.

Transcoding: A technique for content adaptation by modifying the incoming and outgoing HTTP stream.

URL: Universal Resource Locator. URL identifies the address location of the Web pages.

XML: Extensive Markup Language is a W3C standard similar to HTML, but it allows creators to create their own tags.

Xpath: XPath is a language for addressing parts of an XML document. It is used together with XSLT and XPointer.

Xpointer: Xpointer is the XML Pointer Language that defines an addressing scheme for individual parts of an XML document.

XSL: Extensive stylesheet language is a W3C standard that specifies how a program should render XML document data.

Web Site Usability

Louis K. Falk
Youngstown State University, USA

Hy Sockel
Youngstown State University, USA

EVOLUTION

Strictly speaking, the term usability has evolved from one of use to also include design and presentation aspects. A large amount of research has been conducted using this wider definition. These studies include everything from model development (Cunliffe, 2000) to personal self image on Web sites (Dominick, 1999) to the purpose of a Web site (Falk, 2000), and to Web site effectiveness (Briggs & Hollis, 1997). Ultimately, these topics are related to usability and the success a Web site enjoys. The construct of usability covers a range of topics. This paper specifically addresses Web usability from the perspective of how easy a system is to learn, remember and use (Rosen, Purinton & Lloyd, 2004). The system features should emphasize subjective satisfaction, low error rate and high task performance (Calongne, 2001). In this regard, usability is a combination of the underlying (hypermedia) system engine and the contents and structure of the document, and how these two elements fit together (Lu & Yeung, 1998).

USABILITY GOALS

At one time, usability was an afterthought in the computer industry; developers were rewarded for the features of an application, not its usability. Usability was a suppressed and barely tolerated oddity (Nielsen, 2000). Typically, Web usability is interpreted to mean how effective the Web site is at permitting access to its information. Site design should take into account the users' characteristics, experience and context (Badre, 2002; Rau, Liang & Max, 2003; Chen & Sockel, 2001). People rely on their experiences and use semantic models in an attempt to make sense out of the environment. What might seem an easy application for a design team can be awkward and difficult to the end user (Marinilli, 2002). Therefore, it warrants setting usability goals and measuring them before a site goes into production. If a goal is high task performance, a sensible measure might refer to the speed in which the Web pages load, given a particular hardware and software combination (Calongne, 2001). However, if low error rate is the point of interest, click stream data and server logs might be analyzed to isolate the error patterns.

USABILITY ISSUES

Every Web page has an address on the Internet. The more recognizable the address, the easier it is for the user to become brand aware and the more often they might return to the site. The address of the main Web page is typically called the domain name, and appears on the URL address line of the browser. Typically, the Web is used as a marketing tool that allows millions of potential customers to visit a site each day (Hart, Doherty & Ellis-Chadwick, 2000). However, before that can happen, a person needs to find the appropriate Web page. In that regard, many individuals use and depend upon search engines to locate sites of interests. A serious problem is that a Web site's reference may be buried so deep in a search result that it likely will go unnoticed, and hence not be visited. The consequence is not only a usability issue, it also is a visibility/profitability problem. To circumvent this issue, an organization should consider using meaningful Web addresses (URL), descriptive meta tags in the (X)HTML code, key words in titles and paragraphs, and backward links (link referrals) to help enhance placement of a Web site in search results.

While search engines use Web-bots to find the pages on their own, it makes sense to register the site with the search engines so that search criteria can be tailored to the Web site. The Web site's domain name

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

becomes more meaningful to the user if it contains cognitive cues. Studies show that the majority of a Web site's traffic is generated through search engines and directories.

Design Issues

A goal of a Web page should be to quickly deliver quality content in a fashion that does not cause the person to become hopelessly frustrated. In this regard, "*Time* is a very big factor." A general rule of thumb is that a Web page should load in less than 8 seconds; if it takes longer, users typically abort the request and go onto the next page of interest. Based on an average basic bandwidth of the Internet providers, the 8-second rule translates to Web pages that are less than 50,000 bytes. The 50,000 bytes is the total size of the page, including icons, images, links, sound and verbiage. Some users include too many images, which can cause three problems: cognitive disorientation, slow downloads and excessive bandwidth use. Graphics should be used sparingly – only when they add and have a point (Nygaard, 2003).

The primary element in making a Web site usable is its design. Unfortunately, many people are anxious to skip steps and just go for a "product" without considering the "basics." As in the engineering field, the design has to be "defined" up front, along with the goals and objectives of the site. One cannot test quality into a product; quality has to be designed into it. However, designing interfaces is a complex problem, quite different from typical engineering challenges, because it deals with users' behavioral aspects. Inadequate forethought, tight schedules, misconceptions, inappropriate attitudes and priorities – such as "usability is a plus that we cannot afford now" – and lack of professionalism are responsible for many poor sites (Marinilli, 2002).

Like in any other medium, the design should be aesthetically pleasing and balanced. To avoid optical confusion, the background needs to be just that, background. The site should use ample white space so the site does not appear cluttered. A problem that developers face is that they do not know what size monitor the user has, screen size the user is using or the actual display size of the browser. The usability issue includes the fact that each version of each browser type may interpret Web pages slightly differently, with some browser releases not supporting many of the features. This is further complicated because there is a large mix of disparate technologies: different browsers, different versions of software, different machine-based applications. Further, there are a variety of Web-enabled devices besides the standard desktop PC: TVs, cellular phones, watches and PDAs. Each technology is associated with a different set of characteristics that limit its ability to be usable. Most previous systems were developed for viewing on regular-sized monitors. A great deal of developmental effort is needed before the Web sites that were built for traditional monitors can be adapted for successful viewing on portable devices (Huang, 2003).

Hardware and Software Issues

A Web page should be designed so that it can be viewed in the three major resolution (screen) sizes. The original standard resolution displayed 640 rows of 800 pixels each (640 x 800). A popular resolution size is 800 x 600 mode, with 1024 x 768 rapidly gaining popularity and 1600 x 1200 on the horizon. The screen modes present information differently. The smaller the mode, the larger items appear, leaving less room (real estate) for information to be displayed on the screen. Many developers make the mistake of not checking the Web pages in other resolution modes. The mode size has astonishing effects on Web pages. It can cause line shifts, sentences to be broken midline, moved links, and many other irritating manifestations.

Another dilemma that can have an effect on the design is the browser. A browser is an application that retrieves, interprets and displays an online or offline document it in its final Web page format. The most popular browsers are Internet Explorer, Netscape/Mozilla/Firefox and Opera. Different browsers and versions (even within the same vendor) may display items on a Web site differently. In some cases, certain elements and features, such as marquees and blinking, can be viewed with some browsers and not on others. Some sites are designed to use a specific browser, but even these may not work using a previous version of the same browser. To ensure that a site works correctly, the Web site should be checked against the major browsers.

Navigation

Many issues need to be taken into account when creating an easy-to-use Web site. The layout of the screen is central to the user's ability to recognize information. Information must be placed in a logical order, and its physical location should be taken into account. Web page content can be longer and wider than the visible portion on a screen, causing the user to have to scroll down or across to see the rest of the content. Generally speaking, scrolling should be minimized, and avoided on navigation pages, because hyperlinks below the fold (browser bottom border) are less likely to be seen and chosen (Nielsen, 1999). The screen typically is considered to be divided into nine asymmetrical regions (similar to a tic-tac-toe board), with each region associated with its own prominent use characteristics. Typical "European"-style languages read from left to right; therefore, it is generally considered appropriate to put the more important information on the left of the screen so the viewer reads it first before interest withers.

The three-click rule should also be taken into account. Users should be able to access all content on the Web site within three clicks from the home page. The content of the information should also be fresh and up to date (Langer, 2000). Hyperlinks need to be accurate and clearly marked. They should also be placed at the bottom of long pages. Once accessed, these links should change color. Each level in the site should allow the viewer to go back to the previous level and forward to the next. As a viewer gets deeper into the site, a link should be present that allows the user to return to the opening page so that the navigation can begin anew if desired. Nothing is worse than having a user become frustrated because a way to either exit or start again is not present or apparent.

It is very important, in any discussion of hyperlinks, to note that there should be no dead links. It is annoying to go to a site and click on a link and have nothing happen, or to come back with a "404 error" (page not found). It is like reading a newspaper or magazine article and the continuation is not there. Some feel that a link that leads to a page that states "under construction" is equally annoying; if it is not ready, do not post the page.

The Web page itself needs to cater to the needs of the user. Many developers feel that it is extremely important that each Web page contain contact information, or at least link to a page that has contact information. From a user's perspective, it is extremely frustrating to want to place an order and run into problems and not be able contact anyone for assistance. Requests generated from the site need to be monitored. Contact information does not do any good if no one responds. According to "industry standards," all inquiries should be responded to within 48 hours.

A "site map" can be helpful in making a site more user friendly. In its simplest form, a site map lists everything located on the Web site and provides navigational links to get to the information. This is important because it lets the viewer know what is and what is not on the Web site (Krug, 2000).

Color

Color schemes play an important role in usability; they help tie pages together and help with navigation. Color impacts the Web site in many ways: It can add to the value by helping to organize the site, or detract by making it harder to read the Web pages. To help eliminate confusion, page colors and design should be consistent throughout the site. Radically changing a site's "look and feel" may cause the user to question whether they are on the same site. Within a Web page, color can be used as an effective tool to help categorize products. Amazon.com is a great example; the design is the same, but by changing the color code of the Web page depending on the product, it lets the viewer know what category they are shopping in.

A Web site that would otherwise be "perfect" can be totally unusable if the colors are inappropriately chosen. Color contrast is also important; as an example, some sites are not readable (usable) because the background color or the design is as dark as the font color. Contrasting colors need to be used so the viewer can read the information on the site easily. Dark fonts with any light background should work well. Another issue concerning color regards viewers with visual disabilities. There are a few simple rules to follow for viewers with visual disabilities to be able to use the site. Developers should be aware that not all colors are displayed the same across different browsers or machines. The World Wide Web Consortium (W3C) has identified 216

browser-safe colors. Developers should stay away from red and green backgrounds, ensure high contrast between background and foreground colors, and avoid busy background patterns that interfere with reading. To avoid confusion, default hyperlink colors (such as blue for unvisited links and red for links already visited) should be avoided for text.

USABILITY MODELS

Many developers prescribe to the idea that the first step in making a site usable is to think about usability and the information architecture of the site before it is actually developed. Because the success of site is based on the metaphor of how a site will be used, by whom and in what environment, it is essential to define the purpose of the Web site and the expected audience (Rosen, Purinton & Lloyd, 2004). This is an important issue, because it determines the type of information and the breadth, as well as depth. The three basic Web site models (Falk, 2000) are: the Presence Model (often referred to as the "me too model"), Informational Model and E-Commerce Model.

Presence Model

These Web sites are designed to establish a presence on the Web but not really accomplish anything more than "I am on the Web, too." They do not usually contain a lot of information, but they often point to other sites that may. They often are used by individuals to share pictures and such with friends. Organizations have used this model in the past as a promotional tool to show that the company is progressive. This type of site is used mostly by smaller organizations that either don't have the expertise to design a more in-depth site or the manpower to maintain it.

Informational Model

The Web pages in this model are usually heavy with information. These Web pages are set up so the user can get to specific information. A lot of software or computer companies use this model to provide access to Frequently Asked Questions (FAQs) so they do not have to provide traditional support. Organizations that use this model often refer telephone callers to their Web site and consequently miss the opportunity for one-on-one sales.

E-Commerce Model

This model typically employs dynamic Web pages and is designed to create, support and establish sales. There usually is enough information on these sites so that the viewers feel sufficiently comfortable to make a purchase. These sites are run by companies with the expertise to quickly update and maintain online inventories.

FUTURE

The future of Web site usability is changing, not just because of our understanding of how people actually use Web sites; but also because organizations are asking more of their Web presence. New Internet-accessible devices are being introduced, so the earlier semantic metaphor of a "desktop" is often no longer viable. Among the cutting-edge Internet devices are a new breed of portable equipment that enhance issues associated with mobile commerce. The presentation platforms have also grown, and include things such as digital assistance, cell phones, wrist watches, radios and portable marquees. Additionally, software tool vendors are continuously introducing new features and techniques; unfortunately, this often detracts from the organizations' message, rather than add to it.

CONCLUSION

Web site usability is defined as how effective a Web site is at permitting access to that site's information. Certain steps need to be followed to ensure this process. First, the Web site's purpose should be determined before starting the design process. After the purpose is decided, the following elements should be taken into account when designing the Web site so that it is easier to use: load time, aesthetic balance, screen size, browser compatibility, clearly marked hyperlinks, contact information on each page and color schemes. It is important to remember that the definition of the project is critical to the effectiveness

of the site. Further, regardless of the amount of work that went into the process, always check, check and recheck to make sure everything works properly.

REFERENCES

Badre, A. (2002). *Shaping Web usability: Interaction design in context*. New York: Addison-Wesley.

Briggs, R. & Hollis, N. (1997). Advertising on the Web: Is there response before click through? *Journal of Advertising Research*, (2), 33-45.

Calongne, C.M. (2001). Designing for Web site usability. *Journal of Computing Sciences in Colleges,* (3), 39-45.

Chen, K. & Sockel, H. (2001, August 3-5). Enhancing visibility of business Web sites: A study of cyber interactivity. *Proceedings of Americas Conference on Information Systems (AMCIS),* (pp. 547-552).

Cunliffe, D. (2000). Developing usable Web sites – A review and model. *Library Computing,* (3/4), 222-234.

Dominick, J. (1999). Who do you think you are?: Personal home pages and self-presentation on the World Wide Web. *Journalism & Mass Communication Quarterly*, *6*(4), 647-658.

Falk, L. (2000). Creating a winning Web site. *The Public Relations Strategist*, (4), Winter, 37-40.

Hart, C., Doherty, N. & Ellis-Chadwick, F. (2000). Retailer adoption of the Internet: Implications for retail marketing. *European Journal of Marketing,* (8), 954-974.

Huang, A. (2003). An empirical study of corporate Web site usability. *Human Systems Management*, *22*, 23-36.

Krug, S. (2000). *Don't make me think: A common sense approach to Web usability*. Indianapolis, IN: New Riders Publishing.

Langer, M. (2000). *Putting your small business on the Web*. Berkeley, CA: Peachpit Press.

Lu, M.T. & Yeung, W.L. (1998). A framework for effective commercial Web application development. *Internet Research: Electronic Networking Applications and Policy,* (2), 166-173.

Marinilli, M. (2002) The theory behind user interface design. Retrieved January 20, 2004, from *www.developer.com/design/article.php/109 25_ 1545991_1*

Nielsen, J. (1999). User interface directions for the Web. *Communications of the ACM,* (1), 65-72.

Nielson, J. (2000). *Designing Web usability*. Indianapolis, IN: New Riders Publishing.

Nygaard, V. (2003, September 18). Top ten features of a good Web site. Retrieved September 25, 2003, from *http://www.webdesignbits.com/Web_Design/ Web_Page_Design/6*

Rau, P., Liang, P. & Max, S. (2003). Internationalization and localization: Evaluating and testing a Website for Asian users. *Ergonomics,* (1-3), 255-271.

Rosen, D., Purinton, E. & Lloyd, S. (2004). Web site design: Building a cognitive framework. *Journal of Electronic Commerce in Organizations*, (1), 15-28.

KEY TERMS

Browser: A browser is an application that interprets the computer language and presents it in its final Web page format.

Deadlinks: Text or graphics that can be clicked on and then should lead to other information. When accessed, either an error message is returned or the link leads to an under construction page.

Dynamic Web Pages: Web pages whose content vary according to various events (e.g., the characteristics of users, the time the pages are accessed, preference settings, browser capabilities, etc.). An example would be the results of a search via search engine.

Hyperlinks: Text or graphics that can be clicked on to view other information.

Information Architecture: How the Web site's Web pages are organized, labeled and navigated to support user browsing.

Site Map: An overview of all information on the site to help users find information faster

Static Web Pages: The same page content is presented to the user regardless of who they are.

URL (Universal Resource Locator): An Internet address that includes the protocol required to open an online or offline document.

Usability: How effectively site visitors can access a site's information – things enacted to make a Web site easier to use.

Web-Based Learning

James O. Danenberg
Western Michigan University, USA

Kuanchin Chen
Western Michigan University, USA

INTRODUCTION

Web-based learning (a major subcomponent of the broader term "distance learning") is one of the tools with which education is delivered at a distance electronically. There seems to be many definitions, as well as terms, for distance learning, such as "distance education," "distributed learning," "remote education," "online learning" and "Web-based learning," which all may refer to the similar education deliverables. In the mid-1990s, the U.S. Department of Education defined distance education as "education or training courses delivered to remote off-campus location(s) via audio, video or computer technologies" (Lewis, Farris & Levin, 1999). Later in the 1990s, the American Association of University Professors (AAUP) defined distance learning as education in which "the teacher and the student are separated geographically so that face-to-face communication is absent; communication is accomplished instead by one or more technological media, most often electronic" (AAUP, 1999).

Today, a more accurate definition of distance learning might allow for the occasional face-to-face encounter between teacher and student, both physically and electronically, along with the requirements of the teacher and student(s) separated at a distance, where technology is used to bridge that gap. There are three common defining components of Web-based or distance learning:

1. The barrier of place and/or time.
2. The goal of education that is being undertaken.
3. The educational tools to overcome the barriers and accomplish the goals.

Web-based learning also implies that the learning is delivered via modern Internet technology. The objective could be of sharing scarce resources with many geographically dispersed learners, or of providing resources to non-traditional learners, those that would not typically attend a traditional campus. The distinction is important when considering the different requirements of the two groups. Besides overcoming the barriers of place and time, Web-based learning allows for a potential cost savings with specialized courses not typically available on a traditional campus. Students can get training according to their particular learning styles and in a format and time frame suited to their needs and schedule.

THE HISTORY OF WEB-BASED LEARNING

Not so long a go, Web-based learning was nonexistent and distance learning was of limited interest to only a relatively few. Recently, however, due to advances in technology, Web-based learning has become an indispensable resource for educators, students, policymakers and even the corporate world. Although distance learning has been around for 250 years in a variety of formats (including mail, telephone, television, audiotape and videotape), the Internet has made this non-traditional format of education very popular. As the multifaceted environment of the Internet continues to evolve, new forms of electronic multimedia, along with new telecommunications technologies, have reduced the constraints imposed by geographic location.

THE ROLE FOR WEB-BASED LEARNING

Classroom teachers have traditionally relied on many visual cues from their students to enhance the

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

delivery of educational material. The attentive teacher consciously and subconsciously receives and analyzes these visual cues and adjusts the course delivery to meet the needs of the class during any particular lesson. In contrast, the Web-based teacher has few, if any visual cues from the students, and the interaction between teacher and student can be very limited compared to a physical face-to-face contact. Even when tools such as real-time video monitors are in place, the cues from the students are filtered and reaction time altered. It can be difficult to carry on a discussion when spontaneity is distorted by such technical requirements.

Many administrators and policy makers feel the opportunities offered by Web-based learning outweigh the obstacles. The challenges posed by Web-based learning are countered by prospects to:

1. Reach a broader student audience.
2. Handle shortages of certain skilled personnel.
3. Meet the needs of students unable to attend on-campus classes.
4. Involve outside speakers who would otherwise be unavailable.
5. Link students from different social, cultural and economic backgrounds.
6. Cope with a rapidly expanding population.

THE DELIVERY OF WEB-BASED LEARNING

Faculty in Web-based learning environments serve as mentors to their students by assisting with independent learning, including answering questions, directing group activities, providing emotional support, pointing to additional resources and evaluating results. Often, Web-based learning requires a small component of actual face-to-face interaction, so the personal side of education can be preserved and increased.

The pace at which material is delivered is sometimes broken up into modules so students can approach each one differently. The use of modules allows for the best form of communication for a given situation: synchronous or asynchronous. Synchronous communication in Web-based learning utilizes a simultaneous group learning environment, whether on a two-way video feed, on the telephone or face-to-face. Asynchronous communication might be represented in a Web-based learning setting as when teacher and student are communicating by e-mail or by letter. As communication technologies have evolved, Web-based learning teachers and students have found more ways to communicate in a synchronous fashion (Connick, 1999).

WEB-BASED LEARNING TECHNOLOGIES

Technology adoption and effects of technology on the participants are two critical factors that greatly impact the success of a Web-based learning system. This section focuses on three aspects of Web-based learning technologies: strategies for technology adoption, issues involved with technology use and empirical findings.

Technology Adoption Strategies

The most important aspect to consider when determining which of the various instructional technologies to use for Web-based learning has to do with the desired results and the potential of a particular technology to reach those instructional goals and outcomes. The key is to focus on the needs of the learners, the requirements of the content and the constraints faced by the teacher and student. This approach may result in a mix of media, each serving specific needs and fulfilling certain requirements. Reisman, Dear and Edge (2001) suggested a five-strategy model for implementing Web-based learning systems (see Table 1). The applicability of the five strategies largely depends on the goals of teaching pedagogy, technical capabilities of instructors and students, and the overall institution commitment to Web-based learning.

A study by Gibbs, Graves and Bernas (2001) offers a list of evaluation criteria of multimedia instructional courseware for when pre-packaged courseware is the preferred option for Web-based learning. Their list includes information content, information reliability, instructional adequacy, feedback and interactivity, clear and concise language, evidence of effectiveness, instruction planning, support and interface design.

Table 1. Implementation strategies (Reisman et al., 2001)

Strategy	Participant	Process	Connectivity	Student support
1	Individual instructor	Ad hoc development	Personal PC	Class/Office hours
2	Individual instructor	Pre-packaged system	Network	9 to 5
3	Group of instructors	Pre-packaged system	Network	9 to 5
4	Institution	Pre-packaged system	Full Web hosting	24/7
5	Institution	Complete outsource	Full Web hosting	24/7

Issues Involved With Technology Use

Although classroom teaching is frequently used as a benchmark for Web-based learning, technologies involved in Web-based learning may introduce new issues in teaching pedagogy. For example, face-to-face interactions are an assumed aspect in classroom teaching, but they can only be approximated with video conferencing and video chat. These synchronous technologies demand a dedicated server connection and consume large system resources. When a large amount of images are transferred in these communication modes, much of the bandwidth will also be used. Ko and Cheng (2004) developed a system to monitor student exams from a remote site. Snapshot images are transmitted from digital cameras equipped on student computers. The central server can be overloaded with simultaneous connections for image transfers if the frequency of video captures approaches real time.

Asynchronous technologies (such as Web and e-mail) can also pose an increased demand for bandwidth. For example, discussion forums that notify forum moderators with an e-mail when there is a post to the forum can consume more network bandwidth compared with forums without this functionality. The bandwidth and network issues become more of the responsibility of instructors when Reisman et al.'s first three strategies are adopted. Bandwidth becomes an issue because most video files are large. A 5-minute video file with sound and annotation can easily be in the gigabyte range. The sheer size of most video and audio files makes downloading a daunting task. Streaming media technologies, however, make the video or audio files ready to enjoy without having to wait for the whole file to be fully downloaded. However, network congestion can delay delivery of data streams, causing pauses or degradation of video quality.

Empirical Studies of Information Technology in Use

Prepackaged courseware resolves technical issues such as security, consistency and portability, but the effectiveness of a Web-based learning system is also influenced by user perceptions. In applying the technology acceptance model to studying the WebCT courseware, Stoel and Lee (2003) found that prior experience with courseware positively influenced perceived ease of use (PEOU), and PEOU predicts perceived usefulness (PU) and attributes toward the target system. PU also affects attitudes. Both PU and attitudes are predictors of future use intention. Yi and Hwang (2003) found similar relationships from a survey of students using the Blackboard courseware. Additionally, they suggested that perceived enjoyment is an antecedent variable that predicts both PEOU and PU. Simpson and Du (2004) found that the two dimensions of learning style (how a person absorbs information and how a person processes information) impacted students' enjoyment level with courses delivered through WebCT. Frequency of computer use has also been found to affect students' attitudes towards courseware (Basile and D'Aquila, 2002).

Carswell and Venkatesh (2002) assessed student perceptions and reactions to Web-based distance education using two validated theories: the theory of planned behavior (TPB) and the innovation diffusion theory. As predicted by TPB, subjective norm and attitude towards the target system are

positively related to acceptance of Web-based learning systems. However, perceived behavior control was not significant to affect acceptance. The innovation diffusion theory predicts that five variables (relative advantage, ease of use, result demonstrability, visibility, trial-ability and compatibility) can influence acceptance, but only relative advantage and visibility were found to be related to one acceptance aspect – involvement.

KEY PLAYERS OF WEB-BASED LEARNING

Successful Web-based learning programs are established through the dedicated efforts of many individuals and members of the academic organization. Four key players in Web-based learning are the following:

1. Students – The primary role of the student is to learn, despite the obstacles relating to separation of course participants, technological necessities and the requirement that they must have a high level of self-motivation.
2. Faculty – Faculty must develop a working knowledge of the instructional technology, must be effective facilitators despite a limited amount of intimate knowledge of the students, and must work with a diverse group of students with multiple expectations.
3. Facilitators – Acting not only as the instructor's on-site eyes and ears, facilitators are often responsible for the set-up and maintenance of on-site equipment, and act as the intermediary between the student and instructor.
4. Administrators – Influential in planning the educational program, they are also decision makers and consensus builders that fulfill the institution's academic mission.

EFFECTIVENESS OF WEB-BASED LEARNING

The education literature suggests that effectiveness of teaching is strongly linked to learning outcome. Webster and Hackley (1997) examined seven learning outcome variables (student involvement and participation, cognitive engagement, technology self-efficacy, attitudes toward technology, usefulness of technology, attitude toward technology-mediated distance learning and relative advantage of such distance learning) for technology-mediated distance learning. Results suggested that (a) perceived medium richness was related to all outcome variables; (b) instructors' attitudes toward technology affected all learning outcomes except involvement, cognitive engagement and usefulness; (c) interactive teaching styles were related to involvement, engagement, attitudes toward technology and attitudes toward distance learning; and (d) instructor control of technology was related to all outcome variables but involvement and self-efficacy.

Studies have also examined effectiveness using traditional learning outcome variables, such as grade point average, test scores and student self-reports. Buckley (2003) found that there was no difference in student outcomes (as measured in traditional outcome variables) among traditional classroom, Web-enhanced and Web-based courses.

Performance-based outcome variables are one important aspect to measure effectiveness, but the literature also suggests factors other than outcome variables. Psaromiligkos and Retalis (2003) point out that a Web-based learning system consists of human participants, online learning resources and technology infrastructure subsystems. Effectiveness evaluation of such systems should center on issues intercorrelated among these subsystems. Therefore, the following variables for effectiveness evaluation are recommended:

- Contributions of learning resources to the acquisition of knowledge.
- Time spent on tasks using or developing the system.
- Online interactions with peers and instructors.
- Quality of learning resources.
- Learner's profile.
- Preference of learning modes.

Additionally, Wang and Beasley (2002) found students' multimedia preference was not directly related to their performance in Web-based learning systems. However, students with low multimedia preference benefited significantly from the presence of learner control – the degree to which a learner

can direct or control his or her own learning process. Students who prefer a high level of multimedia content were not affected by learner control.

BENEFITS TO PARTICIPANTS

A number of advantages and disadvantages surround Web-based learning. Some of the disadvantages of Web-based learning might be resolved over time, as many online courses being taught today were established in a zeal to create online offerings (Uhlig, 2002). The most obvious advantages for students of Web-based learning include:

1. **Convenience:** courses can be accomplished from any location set up for it.
2. **Flexibility:** students can take courses when they want and at their own pace.
3. **Availability:** since the arrival of the Internet, the number of courses has grown.
4. **Time Savings:** students do not need time to commute to class or to the library.
5. **Career Stability:** working students do not have to relocate or quit their job.
6. **Rich Diversity:** classes are made up of students from many walks of life.

The down side to Web-based learning for students includes:

1. **Commitment:** competing demands from employer, family and friends.
2. **Time Management:** students must be self-motivated and they cannot procrastinate.
3. **Information Management:** students have to actively manage course information.
4. **Technology Savvy:** students must be comfortable with using new technologies.
5. **Acceptability:** there is still a stigma attached to distance learning degrees.

There are rewards for faculty of Web-based learning too, such as:

1. **Rewarding:** fulfilling a desire to work with underprivileged students.
2. **Expectation:** opportunity to work with practicing professionals.
3. **Exhilaration:** thrill of working in new and developing technologies.
4. **Prestige:** the stature of working in a challenging field.
5. **Monetary:** the possibility of extra compensation.

Web-based learning courses have developed a reputation regarding their ability to retain students who enroll. Dropout rates for some online courses hover around 50% (very similar to correspondence courses). However, some characteristics of Web-based courses seem to lead to the retention of those students that enroll:

- Web-based courses labeled 'content-high' include very little instructor-student interaction and little or no student-to-student interaction. These content-rich courses tend to have very high drop-out rates, often between 40% and 50%. Although quite common in first-generation online courses, gradual changes are taking place. Impressive graphics and pretty page designs do little or nothing to reduce the monotony of one-way information.
- Web-based courses labeled 'process-high,' in contrast, involve substantially more interaction and dialog between students and instructors. Often, some course designs encourage this instructor/student dialog as an initial factor, and then reduce the instructor as facilitator once the group is going. Dropout rates in these courses and programs tend to be much lower than with "content-high" courses.

THE FUTURE OF WEB-BASED LEARNING

The rapidly changing technological landscape has made Web-based learning increasingly more accessible and more efficient for many. In the past decade, these advances and the exploding need for lifelong learning have resulted in a significant growth in the number of Web-based learning courses available and the number of students taking these courses. Increasing Internet speeds will bring extraordinary advancements in the integration of graphics, audio

and video into existing Web-based courses. The convergence of the Internet and modern communication technologies is enabling new and exciting ways to learn. Virtual lifelong learning is naturally and seamlessly integrating education into all other human activities worldwide (Stallings, 2000; Rhodes, 2001).

Current trends in telecommunications allude to a future where students and faculty will have essentially unlimited bandwidth to work with. Web-based educational software will increase in capabilities and include features present only in the most sophisticated video games and simulation systems of today. In the near future, the purpose of Web-based educational software may be refocused, adapting from providing a stream of information for the consumption by students to a situation where students require the information in order to succeed, and this environment will be actively sought by students (Downes, 1998).

CONCLUSION

Web-based learning is in place to bring about the first overhaul of the teaching/learning process in several hundred years, since the creation of the university concept or since Gutenberg's printing press. New and developing technologies are revolutionizing how we present instructional content and how we can share high quality, individualized, learning events with students from around the globe.

Several problems exist with Web-based learning that have yet to be overcome. The hurdle of offering hands-on laboratory courses, the loss of instructor control during examinations or for assignments, and the constantly changing technological requirements of remote locations add up to potential barriers to the learning process. Despite these limitations, there is great potential for growth in Web-based learning. The information technology explosion in our global society is creating tremendous challenges and opportunities for educators as we help shape the next generation.

REFERENCES

American Association of University Professors. (1999). Statement on distance education. Retrieved March 30, 2004, from *www.aaup.org/Issues/DistanceEd/intro.htm*

Basile, A., & D'aquila, J.M. (2002). An experimental analysis of computer-mediated instruction and student attitudes in a principles of financial accounting course. *Journal of Education for Business, 77*(3), 137-43.

Buckley, K.M. (2003). Evaluation of classroom-based, Web-enhanced, and Web-based distance learning nutrition courses for undergraduate nursing. *Journal of Nursing Education, 42*(8), 367-370.

Carswell, A.D., & Venkatesh, V. (2002). Learner outcomes in an asynchronous distance education environment. *International Journal of Human-Computer Studies, 56*, 475-494.

Connick, G.P. (1999). *The distance learner's guide.* Upper Saddle River, NJ: Prentice Hall.

Downes, S. (1998). The future of online learning. Retrieved March 30, 2004, from *http://downes.ca/future/*

Gibbs, W., Graves, P.R., & Bernas, R.S. (2001). Evaluation guidelines for multimedia courseware. *Journal of Research on Technology in Education, 34*(1), 2-17.

Ko, C.C., & Cheng, C.D. (2004). Secure Internet examination system based on video monitoring. *Internet Research, 14*(1), 48-61.

Lewis, L., Farris, E., & Levin, D. (1999). *Distance education at postsecondary education institutions: 1997-1998.* Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Psaromiligkos, Y., & Retalis, S. (2003). Re-evaluating the effectiveness of a Web-based learning system: A comparative study. *Journal of Educational Multimedia and Hypermedia, 12*(1), 5-20.

Reisman, S., Dear, R.G., & Edge, D. (2001). Evolution of Web-based distance learning strategies. *International Journal of Educational Management, 15*(5), 245-251.

Rhodes, F.T. (2001). *The creation of the future, an analysis of the modern research university.* Ithaca, NY: Cornell University Press.

Simpson, C., & Du, Y. (2004). Effects of learning styles and class participation on students' enjoyment level in distributed learning environments. *Journal of Education for Library and Information Science, 45*(2), 123-36.

Stallings, D. (2000, Jan). The Virtual University: Legitimized at century's end: Future uncertain for the new millennium. *Journal of Academic Librarianship, 26*(1), 271-280.

Stoel, L., & Lee, K.H. (2003). Modeling the effect of experience on student acceptance of Web-based courseware. *Internet Research, 13*(5), 364-374.

Uhlig, G.E. (2002, Summer). The present and future of distance learning. *Education, 122*(4), 670.

Wang, L.-C.C., & Beasley, W. (2002). Effects of learner control and hypermedia preference on cyber-students performance in a Web-based learning environment. *Journal of Educational Multimedia and Hypermedia, 11*(1), 71-91.

Webster, J., & Hackley, P. (1997). Teaching effectiveness in technology-mediated distance learning. *Academy of Management Journal, 40*(6), 1282-1310.

Yi, M.Y., & Hwang, Y. (2003). Predicting the use of Web-based information systems: self-efficacy, enjoyment, learning goal orientation, and the technology acceptance model. *International Journal of Human-Computer Studies, 59*, 431-449.

KEY TERMS

Asynchronous: Communication between parties in which the interaction does not take place simultaneously.

Computer Assisted Instruction (CAI): Teaching process in which the learning environment is enhanced with the use of a computer.

Computer-Managed Instruction (CMI): Teaching and tracking process in which the learning environment is enhanced with the use of a computer.

Computer-Mediated Education (CME): The developed and still-evolving powerful and sophisticated hypermedia computer tools.

Distance Education: The process of providing instruction at a distance, including the occasional face-to-face encounter between teacher and student, where technology (i.e., voice, video, data and print) is used to bridge that gap.

Distance Learning: The outcome of Web-based or distance education.

Fully Interactive Video: (Two-way interactive video) The interaction between two sites with audio and video.

Synchronous: Communication in which interaction between participants is simultaneous.

Webmetrics

Mario A. Maggioni
Università Cattolica di Milano, Italy

Teodora Erika Uberti
Università Cattolica di Milano, Italy

THE INTERNET: A COMPLEX NETWORK

The Internet is perhaps one of the newest and most powerful media that enables the transmission of digital information and communication across the world, even if there still exist important divides (digital divide) between and within countries in the endowment, access and use of this technology.

To a certain extent, the level and rate of the Internet diffusion reflects its nature, being a complex structure subject to positive network externalities (which are at the basis of the so-called "Metcalfe's law," which states that the value of a network increases with the number of nodes that belong to it: the larger the number of nodes joining a network, the more valuable the network).

In addition, the Internet is a network that evolves dynamically over time; hence, it is important to define its nature, main characteristics and potentialities.

THE INTERNET AND THE WWW

To investigate the nature of the Internet, it is essential to distinguish between the physical infrastructure (which we will call "Internet") and its virtual, graphical and multimedia interface, the World Wide Web (WWW), a service platform which, on January 2004, was made of 4,606,743 pages (Zakon, 2004) and in 2003, its surface was estimated to be equal to 167 terabytes (Lyman & Varian, 2003).

The Internet is a series of connected networks; each of them is composed by a set of Internet hosts and computers connected via traditional or optical cables; while the WWW is constituted by Web pages and Web sites connected by Internet hyperlinks, enabling information and communication to flow from one computer to another. Therefore, the Internet is the physical infrastructure reflecting the technical capability of a given geographical area (i.e., a country, region or city) to enable effective and efficient exchanges of digital information; while the WWW is a virtual space reflecting the ability to create and exchange digital information and contents. Of course, the latter would not exist without the former (Abbate, 1999; Barners- Lee & Fischetti, 1999).

Both the Internet and the WWW are networks, but while the first has a relatively stable infrastructure (because investments to implement and maintain it are rather large and costly, and the subjects involved are a limited number: mostly corporate, governmental or non-governmental organizations), the second changes very rapidly over time (because is cheap and easy to create and maintain a Web site, and the number of people agents involved is huge); therefore, it is very difficult to give a precise and updated description of it.

The most common indicator of the Internet diffusion is the number of Internet hosts, that should reveal the ability of a given geographical area to create digital contents and support the exchange flow of information. Unfortunately, this definition is ambiguous, and its measurement does not entirely capture the actual diffusion of this medium. First, generic Top Level Domains (gTLDs; which account for almost 67% of the total hosts in January 2004) do not reflect any specific geographical location. Second, some country code top-level domains (ccTLDs), even if from a formal perspective, are unambiguously geo-located, display a mismatch between the location of the TLD and the actual source of digital information. For example, .tv domain (acronym for Tuvalu Island) is very diffused among televisions' corporate because of its abbreviation, and the digital contents are not related to this country. Similarly, .nu

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

(acronym for Niue Island) is quite common because of phonetic reasons, but not because Niue inhabitants frequently use the Internet. Third, even if considered jointly with other technological (i.e., number of computers, telephone lines) and economic indicators (GDP per capita), the number of Internet hosts may capture a large share of the Internet infrastructure, but misses mapping the flows of digital information.

Hence, it is crucial to use suitable indicators to map the infrastructure of flows of digital information and contents across the WWW. The number of Web pages and sites reflects the amount of information available on the WWW, but misses the description of the structure of digital information flows, the ability to create digital contents and to attract e-attention (not to mention the crucial issue of the quality of information).

WEBOMETRICS PROCEDURES USING INTERNET HYPERLINKS

A relevant problem in the analysis of the WWW concerns measurement. Almind and Ingwersen (1997), referring to the organizational nature of this service platform – a network of dynamically linked pages – adopted quantitative techniques, derived from bibliometric and infometric procedures, to analyze the structure and the use of information resources available on the WWW. Hence, they introduced the term Webometrics: the bibliometric study of Web pages.

The intuition of these authors was to adapt citations' analysis and quantitative analysis (i.e., impact factors) to the Web space to enable the investigation of Web pages' contents and to rank Web sites according to their use or "value" (calculated through hyperlinks acting as papers' citations); to allow the evaluation of WWW organizational structure; to study net surfers' Web usage and behavior; and finally, to check Web technologies (i.e., retrieval algorithms adopted by different search engines).

The starting point of Webometrics is taking into account the structure of the WWW: a network of Web pages connected through the Internet hyperlinks, strings of text that enable surfing the WWW, whose nature is particularly suitable for this metric analysis.

Although the Internet hyperlinks may refer to different functions (i.e., authorizing, commenting, exemplifying, etc.) (Harrison, 2002), the essential feature for Webometrics procedures is their directionality.

Indeed, Internet hyperlinks are directional, pointing from a page to another one; hence, it is possible to distinguish between the "outgoing" links (i.e., number of hyperlinks pointing to other Web pages importing digital information) and the "incoming" links (i.e., number of links received from other Web pages exporting digital information) (Cioleck, 2002). Second, because these hyperlinks are included into a Web page or site characterized by a domain name, it is easy to assign (under the above mentioned limitations) the ability to offer or demand digital information and contents to a particular player (i.e., country, region, institution or organization). Thus, Internet hyperlinks allow analysts to study the relational structure of the WWW (Cioleck, 2002; Han Woo Park, 2003; Maggioni & Uberti, 2003; Uberti, 2004).

Hence hyperlink based indicators capture the relevance of a Web page or site according to its references (outgoing links) or citations (incoming links) (Almind & Ingwersen, 1997; Björneborn & Ingwersen, 2001; Rousseau, 1997; Thelwall & Smith, 2002). An example of an index calculated in Webometrics analyses is the "Web impact factor," a measure similar to the impact factor calculated in bibliometrics that captures the influence of a site across the whole Web, calculating the number of "sitations," or incoming links, from other sites (Almind & Ingwersen, 1997; Smith, 1999).

Some critics highlight possible drawbacks related to the use of hyperlinks as useful indicators. The first critic refers to the fact that, since inserting an Internet hyperlink in a Web page is a simple and relatively inexpensive action, the informational content of such an indicator is low. The second relates to the fact that different categories of Web sites (i.e., commercial, institutional, academic) may use hyperlinks in totally different ways.

The answer to the first critic highlights that – since the physical space in a computer screen and, above all, the surfer's attention, are limited – there is a "non-monetary" budget constraint that acts as a powerful disciplining mechanism in forcing the Web

designer to limit the number of hyperlinks contained in a given Web page. Many textbooks of Web design show that the number of hyperlinks is a key element in determining the attractiveness of a Web page and that the attractiveness of a page is a non-monotonic function of the number of hyperlinks (Lynch & Horton, 2002). The number of hyperlinks should, in fact, be not too low but also, and more importantly, not too high, since an "empty" as well as "heavy" Web page is considered unpleasant and avoided by a large share of net surfers. This is true for all Web pages, with the exception of search engine Web sites, in which the larger the number of hyperlinks, the better. However, many search engines use the number of incoming hyperlinks (i.e., the number of other Web sites pointing to the targeted one) as a relevance criterion for ranking Web sites on a given topic (i.e., Google adopts a peculiar search algorithm that includes the number of incoming links to rank the search results).

The rationale beyond the presence of an Internet hyperlink are numerous (i.e., functionality, business purposes, semantic or rhetoric reasons), but its presence can be analyzed according to the different purposes of the Web sites. Thus, the second critic may be answered (trading off generality for precision) by limiting the scope of the analysis to the same sort of Web sites (i.e., only commercial or only academic, etc.).

A more extensive use of Internet hyperlinks, according to an economic perspective, may lead to the comparison between a country's position in the structure of exchange of digital information and contents, and in the structure of trade of different goods and services (Brunn & Dodge, 2001; Uberti, 2004); or to the analysis of the inter-regional and intra-regional relationships existing between a number of selected institutions (such as universities, local government, chambers of commerce, etc.) (Maggioni & Uberti, 2004).

The structure of the WWW is also a very interesting topic for topological studies. Several mapping of Internet hyperlinks describe the WWW structure like a scale-free network, where very few Web pages and sites hold a very central position and the rest of the Web is peripheral (Barabasi, 2002).

A map of the WWW in 1999 shows a graph structure similar to a bow-tie with a large center (28% of the total), constituted by the majority of nodes; some nodes (21%) exclusively connected to the center via incoming links; other nodes (21%) connected to the WWW via exclusive outgoing links; some nodes, called tendrils (22%), hanging off the two extremities; and finally, 8% of nodes are completely disconnected (Broder, Kumar, Maghoul et al., 2001).

More recently, some studies emphasized that the Internet hyperlinks infrastructure is not so homogeneous to be represented as a scale-free network, because some typologies of Web pages (i.e., universities and newspaper) have a distribution of hyperlinks that is much more similar to a normal distribution, looking like random networks infrastructure (Pennock, Flake, Lawrence et al., 2002).

Other studies show that the "diameter" of the WWW is very small, compared to its dimension: Two random chosen documents are, on average, 19 clicks away from each other (Albert, Jeong & Barabasi, 1999). This measure reflects a structure of the WWW similar to a "small-world" network (Watts, 1999).

CONCLUSION

Webometrics uses quantitative indexes, derived from bibliometric procedures, to measure different dimensions of the WWW, a service platform constituted by a complex network of Web pages connected through Internet hyperlinks.

Internet hyperlinks constitute a powerful indicator to measure the "Web impact factor" of a Web page and to capture the market's structure of digital contents and information.

Further extensions of this stream of research involve the investigation of actual exchange of information from a computer to another. This would imply to move the focus of the analysis from the mere existence of a hyperlink to its actual "click". In this way, it would be possible to investigate the actual behavior of net surfers and measure the effective and actual exchange of information that flows around the world each day through the Internet.

REFERENCES

Abbate, J. (1999). *Inventing the Internet*. Cambridge: The MIT Press.

Albert, R., Jeong, H., & Barabasi, A.L. (1999). Diameter of the World Wide Web. *Nature, 401*, 130-131.

Almind, T.C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to webometrics. *Journal of Documentation, 53*(4), 404-426.

Barabasi, A.L. (2002). *Linked: The new science of networks*. New York: Perseus.

Barners-Lee, T., & Fischetti, M. (1999). *Weaving the Web*. San Francisco: Harper.

Björneborn, L., & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics, 50*(1), 65-82.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2001). Graph structure of the Web. Retrieved from *www.www9.org/w9cdrom/160/160.html*

Brunn, S., & Dodge, M. (2001). Mapping the "worlds" of the World Wide Web: (Re)structuring global commerce through hyperlinks. *American Behavioral Scientist, 44*(10), 1717-1739.

Cioleck, M.T. (2002). Targets of electronic attention in Asia: who watches whom in the cyberspace? An explanatory study. Retrieved from *www.ciolek.com/PAPERS/electronic-attention2002.html*

Han, W.P. (2003). Hyperlink network analysis: A new method for the study of social structure on the Web. *Connection, 25*(1), 49-61.

Harrison, C. (2002). Hypertext links: Whither thou goest, and why. *First Monday -The Peer-Reviewed Journal on the Internet*. Retrieved from *www.firstmonday.org/issues/issue7_10/harrison/index.html*

Lyman, P., & Varian, H.R. (2003). *How much information?* Retrieved from *www.sims.berkeley.edu/research/projects/how-much-info-2003/*

Lynch, P.J., & Horton, S. (2002). *Web style guide: Basic design principles for creating Web sites* (2nd ed.). Retrieved from *www.Webstyleguide.com/*

Maggioni, M.A., & Uberti, T.E. (2003, February 26-27). Mapping the digital divide: hyperlinks and information flows between European regions. Paper presented at *EU-NESIS Workshop "Regional effects of the new information economy: towards revision of regional disparities indicators,"* Bocconi University, Milan, Italy.

Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J., & Giles, C.L. (2002). Winners don't take all: Characterizing the competition for links on the Web. *Proceedings of the National Academy of Sciences, 99*(8), 5207-5211.

Rousseau, R. (1997). Sitations: an explanatory study. *Cybermetrics, 5*(1). Retrieved from *www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html*

Smith, A. (1999). *ANZAC Webometrics: exploring Australasian Web structures*. Retrieved from *www.csu.edu.au/special/online99/proceedings99/203b.htm*

Thelwall, M., & Smith, A. (2002). Interlinking between Asia-Pacific University Web sites. *Scientometrics, 55*(3), 335–348.

Uberti, T.E. (2004, May 12-16). The architecture of the Internet hyperlinks: a network analysis. Paper presented at *SUNBELT XXIV International Social Network Conference*, Portoro, Slovenia.

Watts, D. (1999). *Small worlds: The dynamics of networks between order and randomness*. Princeton: Princeton University Press.

Zakon, R.H. (2004). *Hobbes' Internet timeline v7.0.* Retrieved from *www.zakon.org/robert/internet/timeline/*

KEY TERMS

Country Code Top Level Domain (ccTLD): The TLD associated to a country, and corresponds to its ISO3166 code. Differently from gTLD, these domains are exclusive of countries.

Digital Divide: This term was first introduced by Clinton's Administration in 1999, analysing the diffusion of computers and the Internet among Americans. Some surveys emphasised the separation between information "haves" and "have nots"

within ethnic groups, and urban/rural population. Later this concept was extended worldwide, distinguishing between countries with a large endowment of ICTs and easy-access conditions to information and countries that have limited endowment and difficult access conditions. Nowadays, the term digital divide refers to differences in the endowment, access and use of new technologies across and within countries.

Domain Name: Any name representing any record that exists within the Domain Name System (DNS), the system that attributes a domain name to an IP address and hence to an Internet host. Three main typologies of top-level domain names (TLD) exist and characterise the ending part of each WWW address: generic top-level domain (gTLD), country code top-level domain (ccTLD) and infrastructure top-level domain.

Generic Top Level Domain (gTLD): TLD reserved regardless of the geographical position. At present, there are the following gTLDs: .aero, .biz, .com, .coop, .info, .int, .museum, .name, .net, .org and .pro. Three peculiar gTLD exist – .edu, .mil and .gov – that are reserved to United States educational, military and governmental institutions or organisations.

HTTP: HyperText Transfer Protocol is the protocol implemented by Tim Barners-Lee to transfer data, messages and information across the WWW.

Infrastructure TLD: The .arpa (address and routing parameter area) domain is reserved exclusively to those who implement the architecture and infrastructure of the Internet.

Internet Host: A domain name that has an Internet Protocol (IP) address record associated with it. This would be any computer system connected to the Internet (via full- or part-time, direct or dial-up connections); that is, nw.com or www.nw.com. (Network Wizards).

Internet Hyperlink (or Hypertext Link): An active link placed in a Web page that allows the net surfer to jump directly from this Web page to another and retrieve information. This dynamic and no-hierarchical idea of linking information was first introduced by Tim Barners-Lee to manage information within a complex and continuously changing environment like CERN. The Internet hyperlinks are directional: outgoing links leaving a Web page and incoming links targeting a Web page.

Metcalfe's Law: Originally defined by Robert Metcalfe in (1993) as follows: The power of the network increases exponentially by the number of computers connected to it. Therefore, every computer added to the network both uses it as a resource while adding resources in a spiral of increasing value and choice.

Web Impact Factor: Similar to the impact factor calculated in bibliometrics, it is a measure of the influence of a site across the entire Web, calculated according to the number of "sitations" in other sites.

Webometrics (or Webmetrics Interneto-metrics or Cybermetrics): First defined by Almind and Ingwersen (1997), it applies bibliometric methodologies and procedures to measure the virtual world of the WWW. This technique refers to the quantitative analysis of the nature, the structures and the properties of Web pages and sites.

URL: Uniform Resource Locator is a system that locates the address of documents and other resources across the WWW.

Wireless Emergency Services

Jun Sun
Texas A&M University, USA

WIRELESS EMERGENCY SERVICES: INTRODUCTION

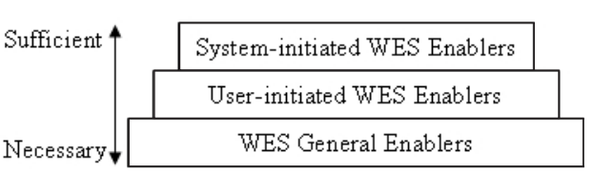
Generally speaking, wireless emergency services can refer to any services that provide immediate help to mobile phone users under emergency conditions. The first widely used WES (wireless emergency services) application is the Wireless Emergency Call Service, which extends the traditional Emergency Call Service (ECS) from fixed-line telephone networks to wireless telephone networks. In the middle of the 1990s, the U.S. Federal Communications Commission (FCC; 1996) issued the order FCC 94-102, requiring wireless carriers to provide the Enhanced 911 Service, the first WECS that delivers emergency calls made from mobile phones as well as caller location information to public-safety answer points (PSAPs). Other countries and regions have planned or implemented similar WECS. For example, the Coordination Group on Access to Location Information by Emergency Services (2002) planned the implementation of E112 service in the European Union.

Though WECS has the capability of pinpointing mobile users, it still responds to user requests like basic ECS, and this type of WES can be denoted as user-initiated WES. However, an emergency is "an unforeseen combination of circumstances or the resulting state that calls for immediate action" (Merriam-Webster Online Dictionary). Thus, emergency events are unforeseeable, but may lead to much worse consequences if no actions are taken immediately. In many cases, people are unaware of or unable to report emergency events, and user-initiated WES cannot help. This calls for another type of WES in which information systems detect the occurrence of emergency events and quickly determine who are (likely to be) involved and what kind of help is necessary. This type of WES can be called system-initiated WES in contrast to user-initiated WES. User-initiated WES and system-initiated WES, together, can provide people with comprehensive protection and minimize loss from mishaps. Demands from a variety of areas, ranging from medical health care, disaster management, to homeland security (e.g., Skinner & Mersham, 2002; Yen, 2004), are pushing WES development forward. This article discusses the latest technologies that make various WES applications possible, and how these applications evolve.

KEY TECHNOLOGICAL ENABLERS

WES applications are based on a variety of technologies: Some serve as infrastructures for both user-initiated and system-initiated WES (i.e., necessary conditions), while others enable each type of WES in specific (i.e., sufficient conditions). Thus, there are three levels of technological enablers: WES general enablers, User-initiated WES enablers, and system-initiated WES enablers (Figure 1). WES general enablers are those infrastructural technologies that support wireless telecommunication and multimedia information processing for both types of WES applications. User-initiated WES enablers include various position-determination technologies that allow wireless carriers to pinpoint mobile users. This positioning capability not only features user-initiated WES, but also makes system-initiated WES possible. System-initiated WES enablers are mainly context-aware technologies that allow WES applications to detect the occurrence of emergency events in user contexts.

Figure 1. The hierarchy of WES enablers



Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

WES General Enablers

In addition to basic voice communications, WES involves the delivery of all kinds of multimedia information over wireless networks, such as textual and graphic messages, user positions, and sensor data. Thus, WES general enablers include new-generation wireless telecommunication technologies that support both voice and data communications, and the latest multimedia and network frameworks that support the integration, processing, and distribution of multimedia information.

New-generation (2.5G and above) wireless telecommunication technologies provide fast and reliable wireless data communications that are essential for WES applications. Most commonly used standards now under the Third-Generation Partnership Project (3GPP, see http://www.3gpp.org), including general packet radio service (GPRS), enhanced data rates for GSM evolution (EDGE)" with (EDGE; GSM - global system for mobile communications), and the universal mobile telecommunication system (UMTS), are packet based. Packet-based wireless networks give users "always-on" capability for data communications, which means users can send and receive information through wireless networks anytime while they are charged based on actual usage rather than connection time. This capability is particularly important for WES applications because users need to keep their WES-enabled mobile phones on standby most of the time for possible emergency events. New-generation wireless telecommunication technologies also provide the necessary bandwidth for WES data communications (e.g., GPRS networks typically transfer data at about 50 Kbps, EDGE networks at 384 Kbps, and UMTS up to 2 Mbps). Reliable and high-speed wireless networks allow the timely transmission of user information (e.g., positions and body conditions) from users to WES systems, and emergency-related messages (e.g., textual notification and evacuation map) from WES systems to users.

The latest multimedia and network frameworks that support WES include the MPEG-21 multimedia framework and the wireless intelligent networking framework. The MPEG-21 framework is a new set of multimedia standards regarding how to adapt digital items related to user delivery contexts (such as user and environmental characteristics) for universal multimedia access (MPEG Requirements Group, 2002; MPEG - Moving Picture Experts Group). Since information involved in WES is mostly multimedia in nature and emergency service delivery is closely related to user contexts, this new multimedia framework is particularly relevant to WES application development. The wireless intelligent networking framework, including the wireless intelligent network (WIN) concept developed by the Telecommunications Industry Association (TIA) and the customized applications for mobile network enhanced logic (CAMEL) concept developed by 3GPP, is about how to deliver intelligent network capabilities to mobile phone users (Christensen, Florack, & Duncan, 2001). Important capabilities for WES include roaming across WES providers (usually wireless carriers), hands-free operation based on voice recognition, and data-service capabilities such as short-message services (SMSs), enhanced messaging services (EMSs), and multimedia messaging services (MMSs; Le Bodic, 2003). These new multimedia and network frameworks support and enhance a variety of important WES functionalities for user convenience and service effectiveness.

User-Initiated WES Enablers

User-initiated WES is featured by its capability to pinpoint mobile users when they make emergency calls. Thus, position-determination technologies are mainly what enable user-initiated WES. There are generally two types of positioning technologies for mobile phones: network based and satellite based (Roth, 2004). Network-based systems use triangulation methods, such as the angle-of-arrival method and time-of-arrival method, to determine user positions as relative to fixed transceivers. Satellite-based systems, usually based on the Global Positioning System (GPS), obtain location information from satellite signal receivers embedded in mobile phones. Satellite-based positioning is usually more precise than network-based positioning.

Position-determination technologies are essential to user-initiated WES because mobile users may be unable to clearly describe where they are when they make emergency calls. It is up to wireless carriers to pinpoint callers so that necessary personnel and equipment can be dispatched immediately. This capability is also necessary for system-initiated

WES applications that, in addition to dispatch services, often need user location information to determine who are (likely to be) involved in certain emergency events and to tailor informational services accordingly.

System-Initiated WES Enablers

In system-initiated WES, information systems detect the occurrence of emergency events and initiate emergency services to users who are (likely to be) involved. Because context-aware computing is about collecting and utilizing user context information in order to provide appropriate services to users (Dey, 2001; Moran & Dourish, 2001), system-initiated WES is one type of context-aware application. In particular, system-initiated WES is oriented toward user emergency contexts, which can be divided into internal emergency contexts and external emergency contexts depending on the physical locus of the contextual elements as related to mobile users (Sun & Poole, 2005). Accordingly, there can be two system-initiated WES applications: personal WES and public WES. Personal WES responds to user internal emergency contexts: mainly abnormal body conditions, such as heart attacks and serious impacts. Public WES responds to user external emergency contexts: mainly natural and human disasters, such as fires, tornadoes, earthquakes, and terrorist attacks. These two newly proposed applications, enabled by context-aware technologies, may become the trend of WES.

Context-aware technologies include several specific technologies, mainly sensor-device technology and sensor-network technology. Sensor-device technology allows the monitoring of the physical environment at different levels with various types of sensor devices (Ristic & Roop, 1994). Sensor-network technology is about how to connect all kinds of sensor devices through wired and/or wireless networks into environment-monitoring systems.

Sensor devices can be generally categorized along two dimensions, mode (contact vs. remote) and locality (fixed vs. mobile), resulting in four coverage levels: individual level, local level, area-wide level, and regional level (Table 1). Wearable sensors, such as those embedded in wristwatches, are mobile contact sensors that monitor the internal contexts, or body conditions, of individual users, such as heart rate and impact force (e.g., Bonato, 2003). Other types of sensors monitor user external contexts, or public environments, at different levels. Fixed contact sensors, such as smoke and fire sensors installed in buildings, monitor the local environment (in the unit of meters). Fixed remote sensors, such as Doppler weather radars, monitor area-wide environments (in the unit of kilometers). Mobile remote sensors, such as surveillance satellites, monitor the regional environment (in the unit of hundreds or thousands of kilometers). These different types of sensor devices enable system-initiated WES applications to detect various emergency events at different levels of user contexts.

User context information collected by various sensor devices needs to be integrated through sensor networks before it can be utilized by WES systems. Recent technological advances in sensor networks, especially wireless sensor networks, greatly enhance the technical feasibility of system-initiated WES applications. Wireless sensor networks enable the collection, selection, and transmission of sensor information through wireless networks for the purpose of environmental monitoring (Culler & Hong, 2004). Both sensor-device technology and sensor-network technology enable system-initiated WES applications to detect all kinds of emergency events and provide timely help to users who are involved.

Table 1. The categorization of sensor devices and their coverage levels

Mode \ Locality	Fixed	Mobile
Contact	Local (Smoke/Fire Sensor)	Individual (Wearable Sensor)
Remote	Area-Wide (Weather Radar)	Regional (Surveillance Satellite)

Note: The examples of sensor devices are given in parentheses.

THE EVOLUTION OF WES APLICATIONS

The advances in WES-enabling technologies have powered WES development, from user-initiated WES applications to more sophisticated system-initiated WES applications. A big picture of WES evolution would be helpful to the understanding of WES developmental trends as well as the relationship between specific WES applications and technological enablers. Because the main user-initiated WES application, WECS, evolves from the traditional ECS, this section begins with a comparison between their architectures. Then it discusses how system-initiated WES applications, including personal WES and public WES, evolve from WECS.

The traditional ECS is based on fixed-line telephone networks that switch emergency calls to PSAPs. At the receipt of emergency calls, PSAP operators inform relevant dispatch agencies of the emergency events and caller positions as described by callers or indicated by telephone numbers. The dispatch agencies, such as police stations, fire departments, and emergency medical services (EMSs), then mobilize personnel and equipment to where the callers are (Figure 2).

WECS is based on new-generation wireless telecommunication networks that support both voice and data communications rather than fixed-line telephone networks. In addition, there are the positioning systems that pinpoint mobile users when they make emergency calls with various position-determination technologies. Wireless carriers deliver emergency calls as well as user location information to PSAPs, which then contact relevant dispatch agencies (Figure 3).

Figure 2. The architecture of the ECS system

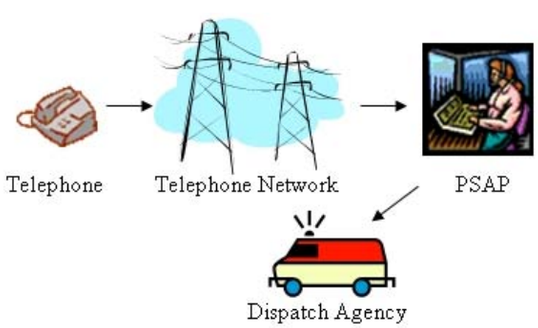
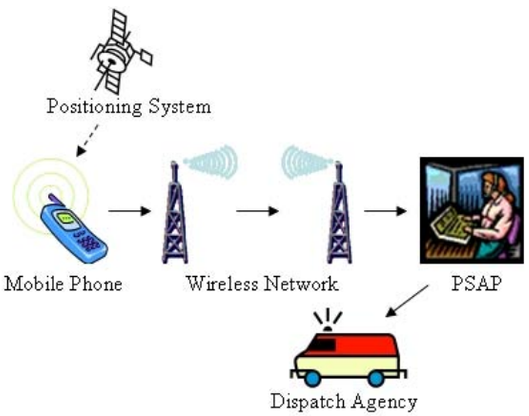


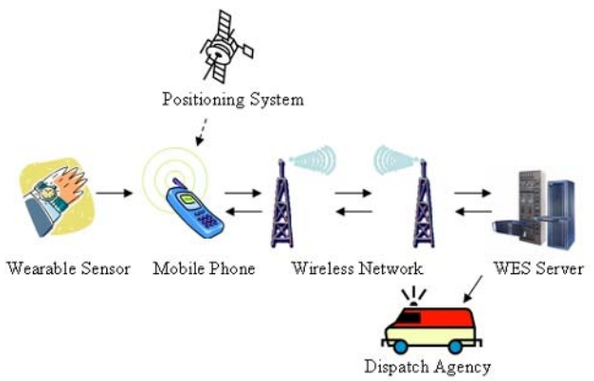
Figure 3. The architecture of the WECS system



As the major user-initiated WES application, WECS is a nonprofit public service mandated by government authorities, but most system-initiated WES applications are value-added mobile-commerce services. Wireless carriers that provide system-initiated WES applications can give subscribers options to register for these services when they activate or change service plans. Successful implementations of system-initiated WES applications can enhance the profitability and competitiveness of wireless carriers. Developed on the infrastructures of WECS, system-initiated WES applications are featured by their context-aware capability powered by sensor-device and sensor-network technologies. Unlike WECS that is mostly voice based, system-initiated WES applications are mostly data based. Personal WES and public WES use high-capacity computers, or WES servers (see Figures 4 and 5), to automatically detect the occurrence of various emergency events, give tailored notification and advice to individual users, and send service requests to relevant dispatch agencies. Thus, in system-initiated WES, central WES servers perform similar functionalities of distributed human-operated PSAPs as in user-initiated WES.

Personal WES uses wearable sensors to monitor the internal contexts of registered users. Abnormal body conditions trigger the sensors to connect with and send sensor information to mobile phones through personal-area networks (PANs) based on Bluetooth™wireless technology. In order to avoid false alarms, mobile phones notify users that abnor-

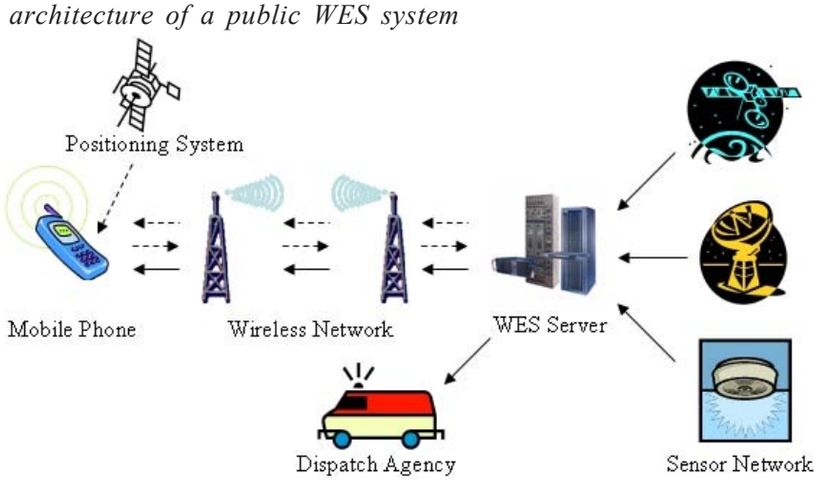
Figure 4. The architecture of a personal WES system



mal conditions are detected, and requests for help would be sent out in a few seconds unless users cancel the requests. If users acquiesce or are unable to respond in serious situations, their mobile phones then send sensor information along with location information to WES servers through wireless telecommunication networks. Based on the information, WES servers find out relevant dispatch agencies, such as local EMSs, and send service requests and user information there. Meanwhile, WES servers call back users and send help-guidance messages to their mobile phones. If users pick up the phone, human WES operators will talk with them. Otherwise, the emergency ringing tones may help people nearby find phone owners quickly. The help-guidance messages tell helpers how to carry out first-aid measures (e.g., cardiopulmonary resuscitation) if necessary (Figure 4).

For instance, if a registered user of personal WES is involved in a car crash, the impact beyond a threshold triggers the g-force (acceleration/deceleration) sensor embedded in his or her wristwatch to connect with the mobile phone through the PAN. The phone rings alarming tones and displays a message: "Excessive impact detected! Requests for help will be sent out in 10 seconds unless you click CANCEL." Suppose that the user is temporarily unconscious: The mobile phone sends impact information and the user position to a WES server. Judging that it is a traffic accident because the impact occurs on a highway, the WES server sends a service request and the user position to the local EMS and police station. Meanwhile, the WES server calls back and sends a message to the mobile phone, such as "Helpers: Please gently move the person to a safe place and check the breathing and circulation…" Compared with typical automatic crash-notification (ACN) applications (Champion et al., 1998), prompting users before sending out help requests solves the issue of false alarms that has annoyed ACN users and providers (Bachman & Preziotti, 2001). Also, calling back and providing help guidance may save critical time for others to find and help users. Of course, personal WES can protect users from many other mishaps, such as heart attacks and electric shocks.

Figure 5. The architecture of a public WES system



Note: The dash lines indicate the polling of user location information before sending detailed messages to users

In WECS and personal WES, individual users or their devices report emergency events. In public WES, however, sensor networks detect emergency events that occur in public environments. WES servers connected to sensor networks process the information about user external contexts at different levels. At the detection of emergency events, WES servers send messages to nearby mobile users through local transceivers notifying them of the events and asking for their permission to pinpoint them. Under the privacy-protection laws in different countries, wireless carriers must obtain consent from mobile users before locating them unless they are directly involved in emergencies (Ackerman, Kempf, & Miki, 2003). Based on user location information, WES servers can determine who are (likely to be) involved in emergency events and tailor informational services accordingly. Compared with wireless emergency-broadcasting services (e.g., Desourdis, Smith, Speights, Dewey, & DiSalvo, 2002), public WES, operated under user permission and personalized with user location information, is less intrusive but more helpful to users. Depending on specific situations, WES servers may also notify local dispatch agencies of the emergency events (Figure 5).

Also take a traffic accident, for example: The sensors carried by passengers or installed along highways sense the crash and send the location and impact information to a WES server through sensor networks (in this case, user wearable sensors are part of the public WES sensor networks). The WES server then polls the mobile users around that area with a message like "A traffic accident just occurred in your area. Click OK to let us know where you are so that we can tell how it may affect you." If a user gives the consent, the WES server pinpoints the user at short intervals with the positioning system in order to calculate his or her speed and direction. If the user is driving toward the accident, the WES server sends him or her a warning message as well as a map showing how to reroute. Meanwhile, it notifies the local police station and EMS of the accident.

CONCLUSION

The latest advances in network and multimedia technologies make it possible to develop ubiquitous, quick-in-response, and context-aware WES applications to protect mobile users from all kinds of mishaps. This article discusses several specific technologies as key WES enablers and how they push WES evolution forward from user-initiated WES applications, mainly WECS, to more sophisticated system-initiated WES applications, including personal WES and public WES.

To successfully implement WES applications, especially system-initiated personal WES and public WES, technical, behavioral, and managerial issues must be addressed. Such issues include quality of service (QoS) regarding timely and reliable information delivery over wireless and sensor networks, user privacy protection, cooperation among wireless carriers for seamless WES coverage, and so on. We hope that this article can enhance further discussions and research in WES application development.

REFERENCES

Ackerman, L., Kempf, J., & Miki, T. (2003). *Wireless location privacy: Law and policy in the U.S., EU and Japan.* Retrieved July 17, 2004, from *http://www.isoc.org/briefings/015/index.shtml*

Bachman, L. R., & Preziotti, G. R. (2001). *Automated collision notification field operational test evaluation report*. Retrieved July 15, 2004, from *http://www-nrd.nhtsa.dot.gov/departments/nrd-12/ACNEvaluation/index.htm*

Bonato, P. (2003). Wearable sensors/systems and their impact on biomedical engineering. *IEEE Engineering in Medicine and Biology Magazine, 22*(3), 18-20.

Champion, H. R., Augenstein, J. S., Cushing, B., Digges, K. H., Hunt, R. C., Larkin, R., et al. (1998). Automatic crash notification. *AirMed, 4*(2), 36-39.

Christensen, G., Florack, P. G., & Duncan, R. (2001). *Wireless intelligent networking.* Boston: Artech House.

Coordination Group on Access to Location Information by Emergency Services. (2002). *Report on implementation issues related to access to location information by emergency services (E112) in the European Union*. Retrieved July 12, 2004, from *http://cgalies.telefiles.de/cgalies_final.pdf*

Culler, D. E., & Hong, W. (2004). Wireless sensor networks. *Communications of the ACM, 47*(6), 30-33.

Desourdis, R. I., Smith, D. R., Speights, W. D., Dewey, R. J., & DiSalvo, J. R. (2002). *Emerging public safety wireless communication systems.* Boston: Artech House.

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal, 5*(1), 4-7.

Federal Communications Commission. (1996). *FCC 94-102: Enhanced 911 emergency calling systems*. Retrieved July 15, 2004, from *http://www.fcc.gov/Bureaus/Wireless/Orders/1996/fcc96264.txt*

Le Bodic, G. (2003). *Mobile messaging technologies and services: SMS, EMS and MMS*. Hoboken, NJ: J. Wiley.

Moran, T. P., & Dourish, P. (2001). Introduction to this special issue on context-aware computing. *Human-Computer Interaction, 16*, 87-95.

MPEG Requirements Group. (2002). *MPEG-21 overview v.5*. Retrieved June 15, 2004, from *http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm*

Ristic, L., & Roop, R. (1994). Sensing the real world. In L. Ristic (Ed.), *Sensor technology and devices* (pp. 1-11). Boston: Artech House.

Roth, J. (2004). Data collection. In J. Schiller & A. Voisard (Eds.), *Location-based services* (pp. 175-205). San Francisco: Morgan Kaufmann Publishers.

Skinner, C., & Mersham, G. (2002). *Disaster management: A guide to issues management and crisis communication*. Cape Town, South Africa: Oxford University Press.

Sun, J., & Poole, M. S. (2005). Context-awareness in mobile commerce: Concepts and applications. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking*. Hershey, PA: Idea Group Inc.

Yen, J. (2004). Emerging technologies for homeland security. *Communications of the ACM, 47*(3), 33-35.

KEY TERMS

Context-Aware Technologies: Technologies that enable the collection, delivery, and utilization of user context information. As key enablers of system-initiated WES, context-aware technologies mainly include sensor-device and sensor-network technologies.

Personal WES: A system-initiated WES application that uses wearable sensors to monitor the internal contexts (i.e., body conditions) of mobile users. At the detection of abnormal conditions, WES systems mobilize local emergency medical services (EMSs) and send help-guidance messages to users' mobile phones.

Position-Determination Technologies: Technologies, either network based or satellite based, that enable wireless carriers to pinpoint mobile phone users.

Public-Safety Answer Points (PSAP): Regional offices where operators receive and screen emergency calls from the public (e.g., 911 in the U.S.A., 112 in Europe) and mobilize various dispatch agencies as needed, such as police, firefighters, and ambulance.

Public WES: A system-initiated WES application that uses large-scale sensor networks to monitor the external contexts (i.e., public environments) of mobile users. At the detection of natural and human disasters, WES systems determine who are (likely to be) involved and tailor informational services with user location information.

System-Initiated WES: A type of WES in which context-aware information systems detect emergency events and provide necessary help to involved users when they are still unaware of or unable to report the events. Depending on whether WES is oriented toward the internal or external user contexts, there are two major system-initiated WES applications: personal WES and public WES.

User-Initiated WES: A type of WES that responds to the service requests made by users when they report emergency events with mobile phones. The main application is Wireless Emergency Call Service (WECS), in which wireless carriers pinpoint

mobile users when they dial emergency numbers and deliver their emergency calls as well as location information to local PSAPs.

Wireless Emergency Services (WES): Wireless-network-based services that respond to emergency events, either reported by people or detected by information systems, with immediate help to those who are involved.

WLAN Security Management

Göran Pulkkis
Arcada Polytechnic, Finland

Kaj J. Grahn
Arcada Polytechnic, Finland

Jonny Karlsson
Arcada Polytechnic, Finland

INTRODUCTION

In a wired local-area network (LAN), the network ports and cables are mostly contained inside a building. Therefore, a hacker must defeat physical security measures, such as security personnel, identity cards, and door locks, to be able to physically access the LAN. However, the penetration capability of electromagnetic waves exposes the data-transmission medium of a wireless LAN (WLAN) to potential intruders (Potter & Fleck, 2003).

WLAN security thus requires reliable protection of data communication between WLAN units and strong access-management mechanisms.

BACKGROUND

Today, WLANs provide acceptable security for most applications, but only if the security requirements are accurately identified and addressed. In addition, active monitoring of WLAN security is needed to detect intrusion attacks, to detect improperly configured security options, and to maintain acceptable security.

A new generation of WLAN management and security tools based on the released 802.11i security standard now offers secure user authentication and protected data communication. These upgrades will quickly replace traditional network- and security-management tools. Therefore, administrating, maintaining, and monitoring WLAN security requires familiarity with the available security technology and corresponding tools and products.

WLAN SECURITY POLICY ISSUES

The rule set in Geier (2002) is an example of a basic WLAN security policy:

- Activate WEP (wired equivalent privacy) at the very least
- Utilize dynamic key-exchange mechanisms
- Ensure NIC (network interface card) and AP (access point) firmware is up to date
- Ensure only authorized people can reset the APs
- Properly install all APs
- Disable APs during nonusage periods
- Assign “strong” passwords to APs
- Do not broadcast service-set identifiers (SSIDs)
- Do not use default SSID names
- Reduce propagation of radio waves outside the facility
- Deploy access controllers
- Implement personal firewalls
- Utilize Internet Protocol Security (IPSec) based virtual private network (VPN) technology on client devices
- Utilize static Internet Protocol (IP) addresses for clients and APs
- Monitor for rogue APs
- Control the deployment of WLANs

These security policy issues should, of course, be updated to reflect recent evolution of WLAN security standards such as the adoptions of the WPA (Wi-Fi protected access) and the IEEE (Institute of Electrical and Electronics Engineers) 802.11i standards.

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

WLAN SECURITY STANDARDS

WLAN standards are introduced by three major standardization organizations: IEEE (IEEE Standards, 2003), Wi-Fi Alliance (Wi-Fi Alliance Portal, 2003), and IETF (Internet Engineering Task Force; IETF Portal, 2003). Most of the standards are issued by IEEE. Wi-Fi Alliance handles the practical implementation of these standards through interoperability testing and certification. IETF is engaged in the evolution of Internet architecture.

Major WLAN security standards:

- IEEE 802.11/WEP
- WPA (based on Draft 3 of IEEE 802.11i)
- IEEE 802.11i (WPA2)

The security in IEEE 802.11 is weak due to the lack of user-authentication mechanisms, and the data-encryption mechanism WEP is a weak implementation of the RC4 (Ron's Code #4) algorithm using static encryption keys (Potter & Fleck, 2003).

WPA, introduced at the end of 2002, was intended to address the WEP vulnerabilities. WPA is based on Draft 3 of IEEE 802.11i (also known as WPA2) to satisfy part of the requirements of the full IEEE 802.11i standard (see Figure 1).

The main features of WPA are:

- The temporal key integrity protocol (TKIP) to provide dynamic and automatically changed encryption keys

Figure 1. A comparison between WPA and 802.11i

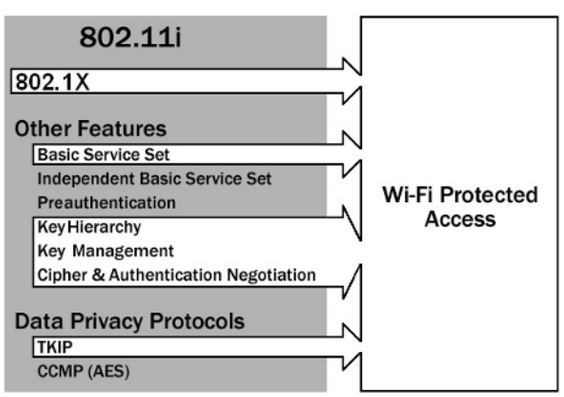


Table 1. Comparison between WEP, WPA, and WPA2

	WEP	WPA	WPA2
Cipher	RC4	RC4	AES
Key Size	40 bits	128 bits encryption 64 bits authentication	128 bits
Key Life	24-bit IV	48-bit IV	48-bit IV
Packet Key	Concatenated	Mixing Function	Not Needed
Data Integrity	CRC-32	Michael	CCM
Header Integrity	None	Michael	CCM
Replay Attack	None	IV Sequence	IV Sequence
Key Management	None	EAP-based	EAP-based

- IEEE 802.1X in conjunction with the extended authentication protocol (EAP) to provide a framework for strong user authentication

The full IEEE 802.11i security standard was ratified by IEEE in June 2004. WPA2 uses the advanced encryption standard (AES) and the encapsulation protocol Cipher-Block Chaining Message Authentication Code Protocol (CCMP) to provide an even stronger data-encryption mechanism than TKIP. WPA2 also supports fast roaming and IBSS (independent basic service set; Edney & Arbaugh, 2003).

A brief comparison between WEP, WPA, and WPA2 is given in Table 1. IV is Initialization Vector, CRC-32 is 32 bit Cyclic Redundancy Check, and CCM is Cipher-Block Chaining Message Authentication Code.

ACCESS MANAGEMENT

Based on IEEE 802.11 Standards

The IEEE 802.11 standard defines open-system and shared-key authentication. SSID and media-access control (MAC) authentication are also commonly used (Potter & Fleck, 2003).

Open-system authentication allows any client to authenticate to a WLAN as long as it passes through a possible MAC address filter. This authentication mechanism is very vulnerable since all authentication packets, including MAC addresses, are transmitted without encryption and MAC addresses are easily "spoofed."

SSIDs are normally broadcasted by WLAN APs. This means that intruders can easily access open-

system WLANs with the use of a mobile device and an NIC. Some AP vendors support disabling SSID broadcasts, but an SSID can still be easily determined by sniffing probe-response frames from an AP.

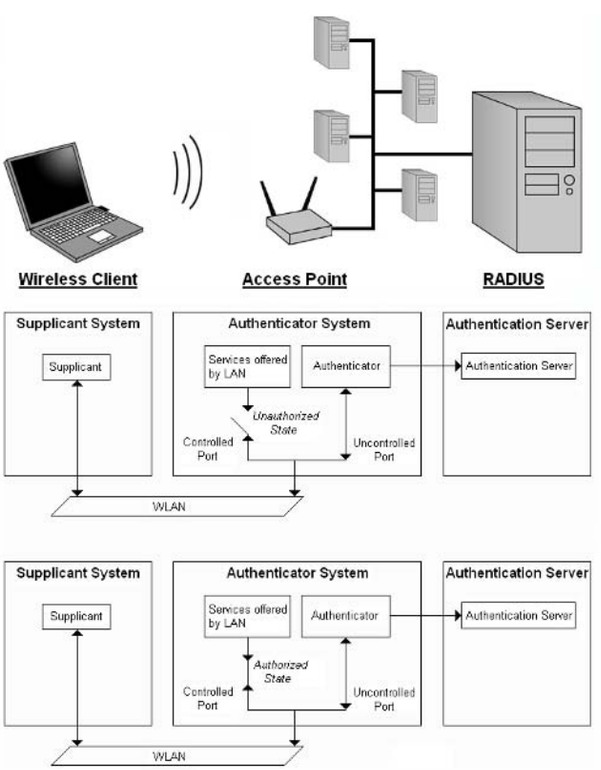
Shared-key authentication is based on static WEP keys that are manually configured into every AP and client in a WLAN. Freely available packages that allow attackers to discover the WEP key can be found in Sourceforge Project Wepcrack (2001).

WPA and IEEE 802.11i provide a more secure shared-key-based authentication mechanism, called preshared key (PSK). WPA in PSK mode is, like WEP, also based on manually entered passwords, but the difference is that the same key is not used for both authentication and data encryption like in WEP (Edney & Arbaugh, 2003).

Based on IEEE 802.1X Standard

IEEE 802.1X is a standard, originally designed for LANs, to address open-network access. This standard has also been found useful for access control in enterprise WLANs. 802.1X has three different components involved: the supplicant (client), authenticator, and authentication server. The supplicant is a user or client who wants to be authenticated. The supplicant accesses the network via the authenticator, which is, in the case of a WLAN, a wireless AP. The authentication server, which is typically a remote authentication dial-in user service (RADIUS) server, works as a back-end server providing authentication service to an authenticator. The authentication server validates the identity and determines, from the credentials provided by the supplicant, whether the supplicant is authorized to access the WLAN or not. The principle of the IEEE 802.1X standard for a WLAN is shown in Figure 2.

Figure 2. IEEE 802.1X access control in a WLAN



802.1X makes sure that only authenticated users are granted access through the controlled port on the wireless access point. Until a user is authenticated, the supplicant can only communicate with the authentication server via EAP-over-LAN (EAPoL) messages, as is shown in Figure 3. During the authentication process, the AP passes authentication messages between the supplicant and the authentication server.

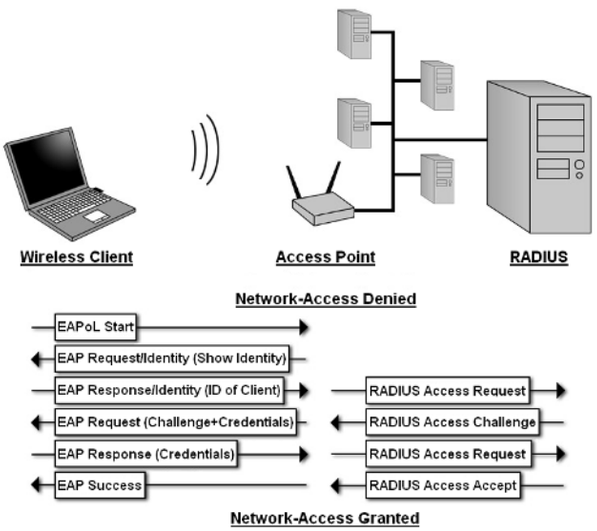
For user authentication, 802.1X utilizes EAP, providing support for several EAP authentication types, such as:

- EAP-MD5 (message digest)
- EAP-LEAP (lightweight, extensible authentication protocol)
- EAP-TLS (transport-layer security)
- EAP-TTLS (tunneled TLS)
- EAP-PEAP (protected EAP)
- EAP-SIM (subscriber-identification module)

EAP-MD5 and EAP-LEAP provide username-password-based user authentication between the supplicant and the authentication server. EAP-MD5 is the least secure version of EAP due to the lack of support for mutual authentication and session-key creation. EAP-LEAP supports mutual authentication and uses a challenge-response exchange.

EAP-TLS, EAP-TTLS, and EAP-PEAP are based on PKI (public-key infrastructure) authentication. EAP-TTLS and EAP-PEAP, however, only use certificate authentication for authenticating the network to the user. User authentication is performed using less complex methods such as

Figure 3. EAP authentication message exchange



username-password. EAP-TLS provides mutual-certificate-based authentication between wireless clients and authentication servers. This means that a X.509-based certificate is required both at the client and authentication server for user and server authentication (Domenech, 2003)

EAP-SIM is an emerging standard allowing users to authenticate to a WLAN using mobile-phone SIM cards and the GSM (Global System for Mobile Communications) mobile-phone authentication network (Radiator EAP-SIM Support, 2004). An EAP-SIM-supported WLAN needs a RADIUS server with a GSM/MAP/SS7 (GSM/Mobile Application Part/Signaling System 7) gateway implemented. During authentication, the RADIUS server establishes a connection with the user's GSM operator through the GSM/MAP/SS7 gateway and retrieves the GSM triplets, which are then used to authenticate the user. EAP-SIM implementations are presented in AMUSE (2004) and EAP-SIM (2004).

Trust Management

The introduction and use of definitions for credentials, trust levels, trust relationships, and security policies are components of trust management. In large-infrastructure WLAN environments, trust management is based on cryptographic techniques such as PKI.

In PKI, public keys are generated, distributed, and certified by CAs (certificate authorities), RAs (registration authorities), and directory services (Housley & Polk, 2001). These entities can be used to establish a hierarchy or chain of trust. Entities that are unknown to each other in a WLAN, such as a user and an authentication server, individually establish a trust relationship with the CA that has issued and signed the user and/or server certificate.

The simplest trust model is the single-point mode with only one CA. All users in this environment can trust each other since all their certificates are issued and signed by the particular CA that they all trust. However, for large environments, this approach has several disadvantages.

- If the central CA is compromised, all certificates will be nullified.
- The central CA could be a congestion point when the number of certificates increases.
- The single entity that runs the central CA should be trusted by all, which might not be achieved in practice.

PKI implementations for large-network environments are based on the hierarchical trust model with one root CA and a number of underlying CAs (Housley & Polk, 2001).

The trust model in EAP-TLS consists of two parts: a client-trusting server and a server-trusting client. At the client, a root CA must be configured to be trusted. By using this root CA, the client is able to validate the authentication server assuming that the configured root CA has signed or is a part of the server certificate's CA chain. Correspondingly, a similar configuration is required at the authentication server in order to make it possible for the server to authenticate the client (Aboba & Simon, 1999).

INTRUSION MANAGEMENT

Intrusion detection is the attempt to prevent unauthorized access to system resources and data, and/or to detect inappropriate, incorrect, or anomalous activity for damage repair later. Intrusion detection will provide an extra layer of protection compared to firewalls and other access-prevention mechanisms.

Intrusion-detection systems (IDSs) detect illegally acting intruders in a computer system. IDSs are categorized in two main groups: host based and network based. A host-based IDS is described as a program process monitoring sensitive activities on a host computer (server). A network-based IDS is located at strategic points in the network in order to detect intrusion activities. A standard IDS architecture consists of four layers (see Figure 4).

- Sensor layer: Sensors gather relevant data from the monitored system.
- Filters: Filters parse the provided data in order to detect possible attack patterns.
- Alert flow: The alert flow is generated by the IDS filters and directed to a monitoring center.
- Monitoring center: The end user monitors and interprets the alert flow.

This IDS architecture can be extended to distributed IDSs where the layers can be implemented on different systems (Northcutt, 2002).

The broadcast nature of wireless networks requires intrusion detection at the data-link layer or at the physical layer when high security is required. Security of wired networks is assured at the network layer or at higher layers; that is, the lower layers are protected physically by the wires. A multidimensional intrusion-detection approach is required since no single method can detect all similar, possible intrusions into a WLAN. Typically, wireless intrusion-detection methods include the following (Lim, Schmoyer, Levine, & Owen, 2003).

- Tracking the MAC address of network adapters trying to associate with the network
- Investigating the relationship between the RTS (request to send) and the CTS (clear to send) frames
- Tracking anomaly data like unsolicited, random responses
- Determining unique signatures for each attack, including characteristics from sequence numbers, control types, destination MACs, SSIDs, organizationally unique identifiers (OUIDs), logical link control (LLC) protocol types, LLC protocol identifiers, and data payload
- Profiling the attack by utilizing rule-based algorithms, expert systems, and artificial neural networks
- Positioning (triangulating) the attacker to determine if the source is a valid user

WLAN intrusion methods are either passive or active. A passive method uses radio-frequency (RF) monitoring, while an active method broadcasts signals to get information about the network or to insert malicious data into the network (Lim et al., 2003). Some common wireless intrusion methods are the following (Potter & Fleck, 2003).

- War driving
- MAC address spoofing
- Rogue access
- Denial-of-service attacks

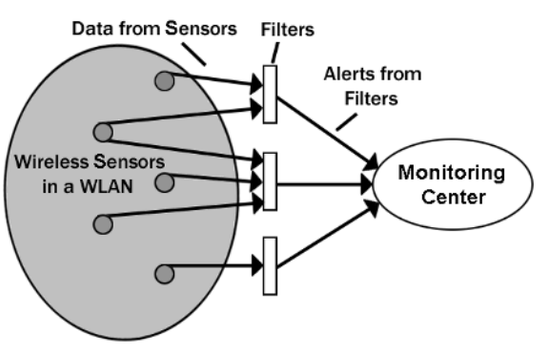
Intrusion attempts based on dictionary attacks on user passwords can, of course, also occur.

SECURITY AWARENESS

The acknowledgement of the following procedures (Simkins, 2004) is considered to be minimal security awareness by a WLAN user.

- Stolen and lost network adapters should be reported.
- The personal installation of APs and tampering of installed APs is strictly forbidden.
- Antivirus software and personal firewalls should be used on WLAN clients.

Figure 4. Basic IDS architecture for a WLAN



The security staff of a WLAN is responsible for the following:

- Auditing the security awareness of users
- Implementing user support and organizing education and training for necessary security awareness

FUTURE TRENDS

New WLAN Authentication Protocols

The recent WLAN security standards WPA and IEEE 802.11i provide trustworthy authentication of WLAN clients and APs, as well as integrity and confidentiality of data communication between authenticated WLAN clients and APs. The same security features can also be obtained by the IPSec protocol for a WLAN with a single AP in an access router.

End-to-end security in client-server applications requires security protocols like VPN, TLS, and SSH (secure shell). However, to prevent the unauthorized use of APs, any end-to-end security solution must be combined with the unambiguous identification of authorized WLAN clients. The required security is obtained by securing the hop between the client and the AP with WPA or IEEE 802.11i in all client-server applications. A drawback of this solution is the high computational load in the WLAN client. Data packets already encrypted with the used end-to-end security protocol must be encrypted once more with the strong encryption algorithm in WPA or IEEE 802.11i.

Two new approaches to WLAN client authentication are presently being proposed. An IETF Internet working group (PANA Working Group Web Page, 2004) is developing a network-layer protocol PANA (protocol for carrying authentication for network access), which can be used also for WLAN client authentication. The other approach is SOLA (statistical one-bit, lightweight authentication), a computationally light protocol for the mutual authentication of a WLAN client and an access point (Johnson, Nilsson, Fu, Wu, Chen, & Huang, 2002).

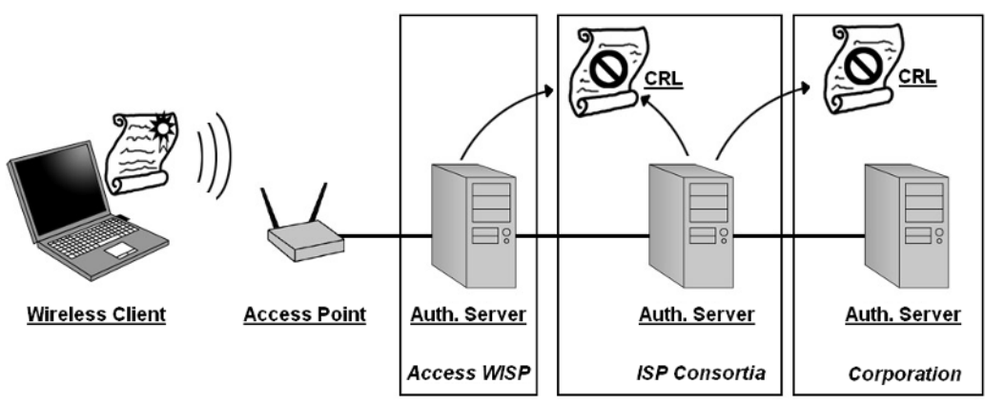
Secure Integration of Different Wireless Technologies

The secure integration of different wireless technologies means secure mobility or secure, seamless roaming in a heterogeneous network environment. Secure, seamless roaming in a heterogeneous network environment including a WLAN and GPRS (General Packet Radio Services)/UMTS (Universal Mobile Telecommunications Service) is briefly discussed in Zivkovic, Lagerberg, and van Bemmel (2004). Mobile IP has been chosen as the solution for session mobility.

Three models for WLAN wireless roaming are presented in VeriSign, Inc. (2004), and Zivkovic et al. (2004).

- **Relationship-based roaming:** WISPs (Wireless Internet service providers) have agree-

Figure 5. Certificate-based WLAN roaming



ments with other ISPs to allow customers to connect to one another's APs, or WISPs or ISPs form a consortia acting as a clearing-house. This model includes storing information about domains, users, routing, management, pricing, and billing.

- **Certificate-based roaming:** A user presenting a certificate is authenticated by the proxy-access WISP. The 802.1X framework with EAP-TLS and authentication servers are deployed. The certificate-revocation list (CRL) issued by the CA is used for verification (see Figure 5).
- **Mobile IP-based roaming:** The WLAN supplicant authenticates to a SP's (service provider's) RADIUS server, not to the WLAN itself. For authentication, the 802.1X/EAP-TLS protocol is used. In order to check user identity and the existence of a roaming agreement, and to forward messages to the RADIUS server a RADIUS proxy is deployed. Once the authentication is verified, the certificate keys are used for data confidentiality. The supplicant obtains a local IP address via DHCP (dynamic host configuration protocol) and registers itself via a mobile IP tunnel at the home agent. Authentication to a GPRS base station involves a SIM card and a home location register (HLR). The client automatically reregisters at the home agent. When the client enters a foreign WLAN, the authentication is done in the same way as in the home WLAN.

Secure Wireless Ad Hoc Networks

Much recent wireless network research focuses on security in mobile ad hoc networks. Basic physical security requires tamper-free network nodes (Stajano & Anderson, 1999). A prerequisite for secure operation is a sufficient level of trust in network nodes (Papadrimitratos & Haas, 2002). Trust and trust relationships depend on network node behavior.

Trust-Management Requirements

Trust management based on manually reconfigurable credentials and network node-interaction rules (Stajano & Anderson, 1999) is possible only in small-scale IBSS networks. In other IBSS networks, trust management is based on cryptographic techniques such as PKI. However, existing trust-management solutions for distributed computer networks cannot be used in the IBSS context because there is no network hierarchy and no central entity in an IBSS network (Papadrimitratos & Haas, 2002).

PKI fault tolerance can be implemented with threshold cryptography (Lysyanskaya & Peikert, 2001). Solutions to distribute certificate-authority functionality across multiple nodes in an IBSS network are presented in Kong, Zerfos, Luo, Lu, and Zhang (2001), Yi and Kravets (2003), and Zhou and Haas (1999). A threshold-signature scheme is proposed in Zhou and Haas.

Secure Routing

The unpredictable and dynamic topology of IBSS networks is a source of routing complexity. Routing in IBSS networks is a rich research field (Giordano, Stojmenovic, & Blazevic, 2003). Routing protocols in computer networks with fixed infrastructures are usually open and unprotected. The Internet draft "Generic Threats to Routing Protocols" (Barbir, Murphy, & Yang, 2003) about routing security was published in December 2003. In an IBSS network, secure routing-protocol operation is a recognized necessity (Papadrimitratos & Haas, 2002).

Secure routing in an IBSS network starts with route-discovery protection by choosing routes satisfying predefined security criteria (Yi & Naldurg, 2001). At node initiation, a route discovery defines the required minimum trust level for nodes participating in the query-reply propagation. Two security extensions of the ad hoc on-demand distance vector (AODV) routing protocol and a routing protocol designed for security in the presence of malicious network nodes, the secure routing protocol (SRP), are described in Papadrimitratos and Haas (2002).

Secure Data Forwarding

Although a correctly discovered route and a secure routing protocol are prerequisites for data forwarding in an IBSS network, there is still no guarantee that the trusted network nodes along a correctly discovered route will indeed relay data as expected. Also, a secure and fault-tolerant data-forwarding scheme like the proposed secure message-transmis-

sion (SMT) protocol is needed (Papadrimitratos & Haas, 2003).

CONCLUSION

Standard organizations such as IEEE, IETF, and the Wi-Fi Alliance have, since the development of WLANs in the late '90s, been working with developing new, reliable security standards to address WLAN vulnerabilities. WEP, introduced in 1999, included serious security flaws due to the use of static encryption keys and the lack of user-authentication mechanisms.

IEEE 802.11i, ratified in June 2004, is expected to address the security flaws in WEP and to eliminate the need of using third-party standards such as IPSec. WPA is today available in most WLAN equipment. WLAN products with full 802.11i support have been available since the fall of 2004. WPA and 802.11i provide reliable access-management mechanisms through the 802.1X standard and use dynamic keys to provide strong data encryption.

Future authentication protocols for 802.11 networks are SOLA and PANA. The enhancement of WLAN security must also include intrusion management as well as security awareness of WLAN service providers and WLAN users.

REFERENCES

Aboba, B., & Simon, D. (1999). *PPP EAP TLS authentication protocol* (RFC 2716). Internet Engineering Task Force. Retrieved September 28, 2004, from *http://www.ietf.org/rfc/rfc2716.txt*

AMUSE. (2004). A mobility trial with UMTS/WLAN seamlessly: Evaluation of the user experience. Retrieved August 26, 2004, from *http://www.surfnet.nl/innovatie/freeband/AMUSE-User-Experience.pdf*

Barbir, A., Murphy, S., & Yang, Y. (2003). *Generic threats to routing protocols* (Internet draft). Internet Engineering Task Force. Retrieved March 13, 2004, from *http://www.ietf.org/internet-drafts/draft-ietf-rpsec-routing-threats-04.txt*

Domenech, A. L. (2003). *Port based authentication for wireless LAN access control.* Master's thesis, Faculty of Electrical Engineering, Technische Universiteit Eindhoven, The Netherlands. Retrieved August 24, 2004, from *http://people.spacelabs.nl/~alex/Port_Based_Authentication_for_Wireless_LAN_Access_Control.pdf*

EAP-SIM. (2004). *Authentication using interlink networks RAD-series RADIUS server and EAP-SIM toolkit.* Retrieved August 26, 2004, from *http://www.interlinknetworks.com/resource/wp5-1.htm*

Edney, J., & Arbaugh, W. A. (2003). *Real 802.11 security: Wi-Fi protected access and 802.11i.* USA and Canada: Addison Wesley Professional.

Geier, J. (2002). *The guts of WLAN security policy* (Tutorial). Retrieved August 29, 2004, from *http://www.wi-fiplanet.com/tutorials/article.php/1499151*

Giordano, S., Stojmenovic, I., & Blazevic, L. (2003). Position based routing algorithms for ad hoc networks: A taxonomy. In X. Cheng, X. Huang, & D.-Z. Du (Eds.), *Ad hoc wireless networking*, 103-136. Boston: Kluwer Academic Publishers.

Housley, R., & Polk, T. (2001). *Planning for PKI: Best practices guide for deploying public key infrastructure.* New York: John Wiley & Sons, Inc.

IEEE standards. (2003). Retrieved December 12, 2003, from *http://standards.ieee.org/*

IETF portal. (2003). Retrieved December 12, 2003, from *http://www.ietf.org/*

Johnson, H., Nilsson, A., Fu, J., Wu, S. F., Chen, A., & Huang, H. (2002). SOLA: A one-bit identity authentication protocol for access control in IEEE 802.11. *Proceedings of GLOBECOM 2002: IEEE Global Telecommunications Conference, 21*(1), 777-781.

Kong, J., Zerfos, P., Luo, H., Lu, S., & Zhang, L. (2001). Providing robust and ubiquitous security support for mobile ad-hoc networks. *Proceedings of the Ninth International Conference on Network Protocols (ICNP'01),* 251-260.

Lim, Y.-L., Schmoyer, T., Levine, J., & Owen, H. L. (2003). Wireless intrusion detection and response. *Proceedings of the 2003 IEEE Workshop on Information Assurance*, 68-75.

Lysyanskaya, A., & Peikert, C. (2001). Adaptive security in the threshold setting: From cryptosystems to signature schemes. In C. Boyd (Ed.), *Lecture notes of computer science: Vol. 2248. Advances in cryptology-ASIACRYPT 2001: Seventh International Conference on the Theory and Application of Cryptology and Information Security,* 331-350. London: Springer-Verlag.

Northcutt, S. (2002). *Network intrusion detection* (3rd ed.). Thousand Oaks, CA: New Riders Publishing.

PANA Working Group Web page. (2004). Retrieved August 31, 2004, from *http://www.ietf.org/html.charters/pana-charter.html*

Papadrimitratos, P., & Haas, Z. J. (2002). Securing mobile ad hoc networks. In M. Ilyas (Ed.), *Handbook of ad hoc wireless networks,* 551-567. Boca Raton, FL: CRC Press.

Papadrimitratos, P., & Haas, Z. J. (2003). Secure message transmission in mobile ad hoc networks. *Elsevier Ad Hoc Networks Journal, 1*(1).

Potter, B., & Fleck, B. (2003). *802.11 security*. O'Reilly & Associates, Inc.

Radiator EAP-SIM support. (2004). Retrieved August 26, 2004, from *http://www.open.com.au/radiator/eap-sim-whitepaper.pdf*

Simkins, R. (2004). *Wireless local area network security.* Retrieved August 16, 2004, from *http://wlan.nat.sdu.dk/WLAN%20Security.htm*

Sourceforge project wepcrack. (2001). Retrieved December 12, 2003, from *http://sourceforge.net/projects/wepcrack*

Stajano, F., & Anderson, R. (1999). The resurrecting duckling: Security issues for ad hoc wireless networks. *Proceedings of the 7th International Workshop on Security Protocols, LNCS.*

VeriSign, Inc. (n.d.). *Secure global roaming for 802.11 WLANs* (White paper). Retrieved March 8, 2004, from *http://research.verisign.com/Papers/VeriSign-WLAN-Security.pdf*

Wi-Fi Alliance portal. (2003). Retrieved December 12, 2003, from *http://www.wi-fi.org/*

Yi, S., & Kravets, R. (2003). MOCA: Mobile certificate authority for wireless ad hoc networks. *Proceedings of Second Annual PKI Research Workshop (PKI03)*, 65-79.

Yi, S., Naldurg, P., & Kravets, R. (2001). *Security-aware ad-hoc routing for wireless networks* (Tech Rep. No. UIUCDCS-R-2001-2241). University of Illinois at Urbana-Champaign, Department of Computer Science.

Zhou, L., & Haas, Z. J. (1999). Securing ad hoc networks. *IEEE Network Magazine, 13*(6), 24-30.

Zivkovic, M., Lagerberg, K., & van Bemmel, J. (2004). *Secure seamless roaming over heterogeneous networks*. Retrieved March 14, 2004, from *http://www.ist-albatross.org/RoamingWhitePaper.pdf*

KEY TERMS

BSS: A basic service set (BSS) is a WLAN architecture consisting of dedicated station computers and a dedicated wireless access point.

Certificate Authority: A certificate authority (CA) is a trusted third party whose purpose is to sign certificates for network entities that it has authenticated using secure means. Other network entities can check the signature to verify that a CA has authenticated the bearer of a certificate.

EAP: The extensible authentication protocol (EAP) is an authentication protocol used with 802.1X to pass authentication-information messages between a suppliant and an authentication server.

ESS: An extended service set (ESS) is a WLAN architecture consisting of dedicated station computers and many dedicated wireless access points.

IBSS: An independent basic service set (IBSS) is a WLAN architecture (also called ad hoc) in which each network unit has both AP and station-computer functionality.

PKI: The public-key infrastructure (PKI) is a system of digital certificates, certificate authorities, and other registration authorities that verify and authenticate the validity of each party involved in an Internet transaction. PKIs are currently evolving and there is no single PKI.

WEP: Wired equivalent privacy (WEP) is a security protocol for wireless local-area networks defined in the 802.11b standard. WEP is designed to provide the same level of security as that of a wired LAN. WEP is used at the two lowest layers of the OSI model.

WPA: Wi-Fi protected access (WPA) is a system to secure wireless networks, created to patch the security of the previous system, WEP. WPA implements part of the IEEE 802.11i standard. In addition to authentication and encryption, WPA also provides improved payload integrity.

Index of Key Terms

Symbols

"Alt" String 307
10 Gigabit Ethernet 54
1G 971
2.5G 971
2G 971
3.5G 971
3G 972
3GPP 972
3GPP2 972
4G 972

A

Acceptable-Use Policy (AUP) 908
access 466, 1000
  control 854, 908
  network 543
  point device 505
  rights management 854
accessibility 232, 1012
action research 943
active learning 135, 784
activity logging 908
ad-hoc network 881, 972
ADL/SCORM ADLNet (Advanced Distributed Learning Network) 935
admission control 323
adoption 6
  factors 473
ADSL 48
advance organizer 709
aesthetic integrity 1012
affective computing 13
affinity communities 1046
AFX 862
agent 614, 916
  attributes 614
  framework 21
agile development methodologies 246
AHS 820
AICC (Aviation Industry CBT [Computer-Based Training] Committee) 935
ambient intelligence 313
anchored learning instructions 94
annotation 695, 1076
antivirus software 569
AOA or Angle of Arrival 636
AOL 1038
Apache 1076
APON or Broadband PON (APON/BPON) 334
application 761
  layer 659
  outsourcing 35
ARPAnet 1038
articulatory synthnetworks 395
ASP (Application Service Providers) 35
assertion of copyright 789
association rules 702
asymmetric DSL (ADSL) 554
asynchronous 1090
  communication 524
  digital subscriber line (ADSL) 328
  distance delivery 224
  transfer mode (ATM) 328, 367, 835, 955
attack
  signature 755
  vs. Intrusion 499
attitude
  toward the ad (Aad) 108
  toward the site (Ast) 109
attribution 789

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

audiographic communication 224
audioteleconferencing 224
auditory icon 347
augmented reality 374
authentication 61, 68, 75, 172, 854, 955, 1000
  system 908
authoring tool 935
authorization 854, 1000
automated port scan 955
automatic facial expression analysis 313
autonomous software agent 21
autopoietic system 868
availability 403
Avatar 426, 592, 840

B

balanced scorecard (BSC) 584
bandwidth 48, 172, 293, 328, 422, 554, 659
banner 88
base station
  controller (BSC) 150
  transceiver (BTS) 150
baseband 203
Baud 422
BBA 862
behavioral biometric 62
behaviour aggregates (BA) 874
benchmark audiovisual affect database 13
benefit function 776
BIFS 862
binary image 769
biometric(s) 6162, , 68, 75, 410
  encryption 68
B-ISDN (Broadband Integrated-Services Data Network) 367
bit
  depth 276
  rate 777
Block Data Hiding Method (BDHM) 769
Bluetooth 172, 301, 651, 972, 984
Bookshelf 723
Boolean Query 122
brand 101
B-Rep 1032
broadband 163, 334, 422, 436, 554, 659
  access 48, 81
  networks 29
  transmission 81
  wireless access 81
broadcast TV services 686
broadcast 517
browse 282
browser 423, 1082
BS 323
BS7799-2:2002 848
BSS 1112
burst(s) 806
  assembly 805
  offset 806
business
  games 537
  performance 416
  strategy 416
  system planning (BSP) 902
  -continuity planning 403

C

cable access 81
cache 736
CAD (Computer Aided Design) 427, 1032
CAP 48
caption 307
carrier 203
  -neutral collocation facilities 163
case study 936
CBT 741
CDMA 972
CDMA-2000 972
CDPD 972
cellular
  communication 643
  network 659
  value-added service categories 584
central route to persuasion 269
CERN 868
CERT/CC© 499
certificate authority 854, 1112
CGI 1077
characteristics of virtual communities 1046
chat 950
chief information security officer 902
CIF (Common Interface Format) 777
circuit
  switched 293
  switching 651
client 761
  -server model 614, 881
clip 736

cluster 695
clustering 828
CMC 94
CMCS 524
CMT 94
coaction field 374
Codec 276
coefficient of determination 784
cognitive theory 741
coherence bandwidth 323
collaborative
  learning 381
collaborative thinking 531
collective action 902
colocation 524
color histogram 41
commercial off-the-shelf (COTS) applications 178
common costs 150
community
  network/civic network (CN) 943
  of practice (CoP) 531, 943, 1063
  -centered development (CCD) 943
competitive thinking 531
compression 979
computed
  radiography (CR) 828
  tomography (CT) 828
computer
  assisted instruction (CAI) 1090
  -managed instruction (CMI) 1090
  -mediated communication (CMC) 241, 730, 840, 950
  -mediated communication/computer-supported collaborative work 1063
  -mediated education (CME) 1090
concatenative synthesis 962
concept map 709
concurrent models with an object-oriented approach 676
conditional access (CA) services 686
condominium fibre 163
confidence 702
confidentiality 75, 403, 1000
congestion control 323
conjoint analysis 6
connected modles with an object oriented approach 676
constraints 676
constructivist perspective 135
consumer context 129
contact 466
content
  analysis 592
  management 115
  repurposing 115
  -based access 41
  -based retrieval 22, 715
content-driven services 686
context
  of use 1019
  -aware technologies 1102
  -sensitive HCI 13
control
  measures 894
  objectives for information and related technology (COBIT) 848
  packet (or Bburst header packet or setup message) 806
controlled vocabulary 307
controlling 510
controls 403
convergence 577
  factor 579
  index 578
COO cell or origin 636
cookie(s) 479, 485
copper line 554
copyleft 789, 797
copyright 561, 790
correlation coefficient 784
country code top level domain (ccTLD) 1094
courseware 936
cover work 389
crackers 410
CRC (Cyclic Redundancy Check) 323
CRISP-DM (Cross-Industry Standard Process for Data Mining) 702
critical digital-mass index 578
CRM (Customer Relationship Management) 35, 578
cross-correlation 323
cryptography 389, 410
CSCL 381
CSCW 1026
CSG 1032
CSMA 607
current cost 150
customer lifetime value (CLV) 676

CVE 916
cyclic graph 1032

D

dark fibre 164
data
  hiding 769
  integrity 75, 1000
  mining 307, 395, 479, 702
  tampering 955
database management system (DBMS) 276
dataveillance 702
DBA 289
deadlinks 1082
decision rule 695
decoder 436
Deixis 307
denial-of-service (DOS) attacks 908
dense-wavelength division multiplexing (DWDM) 164
developing countries 453
device 761
dialect 466
dialogue 858
differential correction 351
differentiated services (DiffServ) 874
diffusion 6
digital
  certificate 485, 854
  divide 1095
  filter 195
  imaging and communications in medicine (DICOM) 403, 828
  rights management (DRM) 812
  signal 195
    processing 195
  signature 854
  subscriber line (DSL) 81, 328, 423
  television (DTV) 203, 517, 686
  video broadcasting (DVB) 203, 686
  watermark 211
  watermarking 389
directed instruction 135
directional relation 264
dispersed/distributed teams 232
distal context 129
distance
  education 1090
  learning 886, 1090
distanced leadership 232
distributed
  environment 446, 665
  models with an object-oriented approach 676
DMT 48
DoCoMo 341
domain name 1095
DSL (Digital Subscriber Loop) 55, 554
DSSS 972
dynamic Web pages 1082

E

EAP 1112
Earcon 347
economic risks 894
EDGE (Enhanced Data for GSM Evolution) 294, 972, 367
EDI (Electronic Data Interchange) 341
EFM Fiber 334
Egress Filtering 755
e-journal 723
elaboration likelihood model 270
e-learning 55, 276, 493, 741, 820
  (distance learning, Web-based learning, online learning) 1054
  process 936
electronic
  catalogs 246
  commerce (e-commerce or EC) 101, 109, 246, 282, 628, 984, 1012
  customer relationship management (eCRM) 676
  data interchange 282
  fund transfers 282
  media 357
elementary entity 695
e-marketplace 894
emergent behavior 21
emotional intelligence 13, 524
EMS 301
encapsulation 659
encoding 276
encrypting 769
encryption 172, 955
enhanced TV 686
enrolment 68
entertainment 109
  service 584
environmental Factors 473

Ephemeris Data Parameters 156
Eriksson-Penker Process Diagram 677
ERP (Enterprise Resource Planning)35
ESS 1112
Ethernet frame 289
Ethernet PON (EPON) 334
European Telecommunications Standards Institute (ETSI) 203
ex ante predictors of student performance 784
experiential learning 537
experiment 592
expertsystems 395
exploit 499
external support 510

F

face-based interface 313
fair information practices (FIP) 479
false
  acceptance 68
    rate 62
  rejection 68
    rate 62
FBA 862
feature extraction 122
FHSS 972
fiber optic cable 328
  -to-the-home (FTTH) 334
file formats 518
financia service 584
finite elements method 427
FIPA 916
FIR filter 195
firewall 410, 755, 956, 1000
flash 886
flow 874
forking 798
formant 962
  synthesis 963
forward error correction 48
forwarding equivalence class (FEC) 874
forward-looking long-run incremental costs 543
fragile watermarking 218
rragmentation 499
frame rate 276
  relay 956
  size 276
free riding 812
freely available 790
frequency-division multiplexing 48
front and back region 357
FTP 381, 868, 1038
FTTx 164
fully interactive video 1090

G

general public license (GPL) 561, 790
generally accepted system security principles (GASSP) 848
generic Top Level Domain (gTLD) 1095
genetic algorithm 29
geostationary satellite (GEO) 351
Gigabit PON (GPON) 334
GIS or geographical information systems 636
global
  popularity 736
  system for mobile communication (GSM) 150
  systems mobile (GSM) 991
globalization 950
globalize 101
GNU 561
GOMS 1019
Google AdWords 270
GPRS 294, 972
GPS (Global Positioning System) 607, 1038
grayscale image 203
GSM (Global System Mobile) 294, 636, 651
GSMC 972
guidelines 518
  for the management of IT security (GMITS) 848

H

H.323 55
hackers 410
haptics 840
HASP (High-Altitude Stratosphere Platform) 367
heterogeneous agents 21
heurictic rule 21
hierarchical conjoint analysis 7
high
  -context culture 241
  -definition television (HDTV) 203
historical costs 150
history retrieval 446
home page (portal) 1054
homogeneous agents 21

host-multicast 881
  routing protocol 881
HTML 307, 979, 1077
HTTP 1077, 1095
human
  capital 525
  -computer interaction (HCI) 13, 537, 1012
  -computer interface 13
HyperClass 374
  HyperLecture, HyperSeminar, HyperTutorial 1068
hyperlinks 1082
hypermedia 178, 460
HyperReality 374
HyperSchool, HyperCollege, HyperUniversity 1068
hypertext 178, 460
HyperWorld 374

I

IBSS 1112
ICAT 916
ICMP Message (Internet Control Message Protocol) 755
ICR (intelligent call routing) 88
identification 75, 232
idiomatic expression 466
IDS (intrusion-detection system) 755
IEEE 802.1x Standard 908
IEEE LTSC (Learning Technologies Standards Committee) 936
IIR Filter 195
IM (instant messaging) 88, 991
image
  classification 702
  clustering 702
  copyrights 212
  indexing 264, 703
  mining 703
  retrieval 703
IMARC 991
i-Mode 991, 301
imperceptible 212
implementation risks 894
impulse response 195
incentive regulation 543
incremental
  cost 150
  development 179
index 715
individual
  critical mass 7
  learning 381
info pyramid 1006
information
  and communications technology (ICT) 453
  architecture 1082
  delivery theory 741
  hiding 389
  requirement elicitation (IRE) 129
  retrieval (IR) 715
  security 902, 1000
    management 395
    policy 902
  service 584
  technology 840
  visualization 1026
  -security-management system (ISMS) 403
informativeness 109
informed embedding 218
innovation factors 473
input debugging 755
inquiry style 1054
installed base 7
instructivist perspective 135
integrated services (IntServ) 874
integrated services digital network (ISDN) 423, 460
integrity 403
intellectual
  capital 1063
  property (IP) 790
intelligent
  agent 614
    technology 395
  algorithms 30
  software agent 709
interaction 246
  channel (IC) 203
interactive
  advertising 109
  digital multimedia 179
  learning 460
  multimedia method (IMM) 460
  multimedia techniques 453
  services 686
  system(s) 246, 1012
  television 436
interactivity 135, 436, 886
intercultural communication competence 241

interface 537, 709, 979
intermedia Transcoding 1006
internal Context 129
international
  outsourcing 798
  standards organisation (ISO) 848
  virtual office (IVO) 466
internationalization 950
internationalize 101
Internet 423
  adoption 473
  host 1095
  hyperlink (or hypertext link) 1095
  protocol (IP) 282, 328
  service provider (ISP) 423, 600, 755
  streaming 277
interoperability 493
interpolation 427
interpretation 950
interstitial 88
intonation 963
intranet 282
intrusion
  detection 755, 902
  prevention 755
  tracking 755
investment 505
ionosphere 157
IP 172, 600
  multicasts 55, 881
  traceback 755
  -multicast routing protocol 881
irrevocability 68
irritation 109
ISDN 956
ISO 1020
ISO 17799 902
ISO 9241 1012
ISO Norm 1032
ISO/IEC 17799 848
isochronous 703
IT
  alignment 416, 510
  implementation success 416
  strategy 416
ITS 820

J

JADE 916
JITAITS 1068

K

Kbps 294
Kerberos 908
key frame(s) 115, 703
killer app 812
knowledge 358
  building 95
  discovery in databases 703
  management 531, 1054

L

LAAS 156
label 874
  swapping 874
LAN 600, 835
last mile 543
LDT 363
leader-led e-learning 741
leading 510
learner model 820
learning
  community 95, 1063
  management Systems (LMS) 936
  network 531
  object(s) 256, 709, 820, 886
  or information ecology 95
  style 709, 886
  -object metadata (LOM) 256, 493
least significant bit (LSB) 389
licensing domain 790
light weight IDS 499
limit conjoint analysis 7
line 544
linear predictive coding 963
lip reading 313
LITEE (Laboratory for Innovative Technology and Engineering Education) 723
live update 569
LLID 289
LM or the location manager 636
LMS 820
LO 741
local
  access 544
  area radio network 651
  loop 544, 554
  network 544

popularity 736
telecommunications market 544
-area network (LAN) 164
localization 628, 950
localize 101
logging 499
LOM 820
lossy compression 218
low
earth orbit (LEO) 351
-context culture 241
-definition television (LDTV) 203
Lurker 841

M

MAC (Medium Access Control) 367, 607
machine
interactivity 730
learning 313, 499
translation (MT) 950
vision 314
macromedia flash 979
magnitude response 195
main distribution frame (MDF) 554
MAN 600
management support 473, 510
MANET 607
MAS 916
mass-spring system 427
m-business 628
media search engine 307
mediated interaction 358
message
sensation value 270
-based service 584
metacognition 95
metadata 307, 923
Metcalfe's Law 1095
metric relation 264
metropolitan-area network (MAN) 164
micro
Browsers 1020
/macro payment 621
microbrowser 115
microelectromechanical systems (MEMSs) 984
middleware 55
mixed media 723
m-learning 820, 1026
MLSE 3230
MMS 301
MMSE 323
MMT 741
mobile
agent 614
commerce (m-commerce) 301, 621, 505, 628, 643
security 621
computing 628
data services (MDS) 991
device(s) 628, 644, 709
HCI 1026
IPv6 294
learning (m-learning) 709
location based services 636
multimedia 644
payment 621
phones 972
switching center (MSC) 150
mobility 628
modality 347
mode 347
modeling 446, 664
modem 48
morphing virus/polymorphicvirus 569
movie 723
m-payment 628
MP3 703
MPCP 289
MPEG 48, 703, 862, 1006
MPEG-1 1006
MPEG-2 203, 1006
MPEG-21 1006
MPEG-4 81, 703, 777, 1007
MPEG-7 703, 1007
MPLS 1007
MPQoS (Mean Perceived Quality of Service) 777
m-security 628
MUD 592
MUD object oriented (MOO) 423
MUD/MOO 841
multi-agent system (MAS) 21, 614
multicast 460
transport protocol 881
multimedia 136, 179, 665, 709, 723, 979
communication 686
data mining 703
database(s) 41, 277, 715
document 264, 715
home platform (MHP) 203

information retrieval (system) 715
service 436
transmission 30
multimodal 347
multimodal (natural) HCI 14
multi-modality 715
multipath propagation 323
multiple channel per carrier (MCPC) 453
multiple service operators (MSOs) 328
Multiplex 651
multiuser dimension/multiuser domain (MUD) 423
municipal fibre network 164
MVNO 301

N

namespace 924
narrowband 335
national culture 241
natural monopoly 544
near-far effect 323
need
context 129
-to-know access policy 902
Netiquette 525
netnography 592, 1046
network 246, 423, 761, 1000
architecture 806
effects 7
interface card (NIC) 505
intrusion 755
layer 659
service providers 35
networked mobile digital devices (NMD) 991
neural network 212
non-repudiation 68, 75, 1000
NPV 505

O

object recognition 703
objective measurement of PQoS 777
OELE 136
OLT 289
one-way reservation schemes 806
online 666
community 944
community of practice 944
discussion board 784
ontologies 493
ontology 924
ONU 289
open
knowledge initiative (OKI) 256
source 790
licensing model 790
software 499, 561, 790, 798
OpenGL ES 862
operationally critical threat asset and vulnerability evaluation (OCTAVE) 848
operator 761
optical
attenuation 835
cross-connect (OXC) 806
line terminal (OLT) 335
network terminal (ONT) 335
window 835
-access network (OAN) 164
opt-in/opt-out 479
option 506
organisational factors 474
organising 510
organization for economic co-operation and development (OECD) guidelines 479
organizational culture 241
outcome-relevant involvement 270
out-of-the-box experience (OoBE) 991
overlay network 813
ownership 790
ownership by contract 790

P

packet 172
switched 294
switching 651
pan 115
participatory design 944
partition 769
passive optical network (PON) 335
patch 1000
PBX 600
PDA (personal digital assistant) 115, 636, 991, 1026, 1037
peer review method 381
Peers 868
peer-to-peer (P2P) network 607, 881
per-hop-behaviour (PHB) 874
peripheral
cue 270

complexity 270
route to persuasion 270
personal
area radio network 651
WES 1102
Petri Nets 446, 485
phase response 196
phishing 410, 485
phonetics 963
phonology 963
photonic networks 972
physical/physiological biometric 62
picture-archiving and -communication system (PACS) 403, 828
Piecemeal Sesign 902
PKI 301, 1113
planning 510
PON 289
popularity 736
port 410
position-determination technologies 1102
power
distance 466
law 868
PQoS (Perceived Quality of Service) 777
principal
-component analysis 427
of least privilege access policy 902
privacy 479
impact assessments (PIA) 485
protection 75
seals 479, 485
private network 956
privilege 854
delegation 854
probability value 784
profile 628
programming language 518
progressive encoding 277
project learning or project works 95
proprietary software 798
prosody 963
protocol 172, 423, 659
data unit 659
proximate context 129
proxy 736, 1077
cache 736
pseudo random noise (PRN) 157
pseudocommunity 944
PST 741
PSTN 600, 972
public
domain 790
land mobile network (PLMN) 150
network 956
switched telephony network (PSTN) 554
WES 1102
publicly available 790
public-safety answer points (PSAP) 1102
PVC 600

Q

QAM 48
quadrature
amplitude modulation (QAM) 203
phase-shift keying (QPSK) 203
quality
degradation 777
of service (QoS) 30, 55, 150, 600, 703, 806, 874
quantization index modulation (QIM) watermarking 218
query by example 41, 122, 716

R

radio-frequency identification (RFID) 984
radius 908
rapid application development (RAD) 179
RDF 1077
reach 730
Real Time 703
record 761
reciprocal communication 730
redundant arrays of inexpensive disks (RAID) 828
reflective practice 136
Regional Bell Operating Company (RBOC) 328
regression 784
relationship risks 894
relevance feedback 41, 122
reliability 881
remote
collaboration 374
user 956
representative region 695
re-purposeable learning objects 886
resource
description framework (RDF) 493
reservation protocol (RSVP) 874

retrieval stage 41
reversible watermarking 218
RFID 1026
ripping 813
risk analysis 902
risks 894
  analysis 956
robust 212
  watermarking 218
router 423, 659
routing
  protocol 607
  table by profile (RTP) 736
RPC 614
RRM (Radio Resource Management) 323

S

salient object 264
satellite 352
  constellation 352
scanning (can be scheduled or batch) 569
scavenging 956
SCU 806
search engine 101, 282, 868
security
  category 902
  clearance 902
  evaluation 956
  event management (SEM) 395
  laboratory 908
  management 499
  policy 395
self
  paced e-learning 741
  -organizing group 1063
semantic
  gap 122
  Web 256, 493
  -based representation 695
semi-fragile watermarking 218
servant 813
server 761
  -based computing (or thin-client technology) 35
service
  access point 659
  level agreement (SLA) 35, 874
session layer 659
set top box (STB) 436, 518, 686
severe acute respiratory syndrome (SARS) 828
shareable content object reference model (SCORM) 256, 493
shareware 790
short
  messaging service (SMS) 992
  -term memory 709
shot 703
signal 196
silent commerce (s-commerce) 984
similarity
  matching 41
  measure 122
single channel per carrier (SCPC) 453
site
  map 1083
  organization 109
  specification 101
situated learning 95
SLA 600
small firm 416, 474, 510
smartphone 992
SNR 48
SOAP (Simple Object-Access Protocol) 88
social
  community 95
  learning 1063
  presence 232
  realities 358
  -construction 358
Socratic dialogue 784
software
  agent(s) 21, 916
  project 381
SONET 55
soundscape 347
source code 798
space constraints 741
spam 485
spatial-temporal activity level 777
spatio-temporal relation 264
splitter 48
spoofing 68, 485, 813
spread spectrum 972
  (SS) watermarking 218
stable filter 196
stages of the virtual community life cycle 1046
standard
  positioning service (SPS) 157
  setting 790

standards 506
start-up latency 736
statement of applicability 403
static
  graphics 979
  Web pages 1083
stealth virus 569
steganography 389, 769
stego
  -object 389
stereotypes 677
stickiness 730
storage-area network (SAN) 828
stovepipe systems 924
stream 874
streaming 115
student participation in the online discussion 784
subscriber
  identity module (SIM) 621, 992
  line 544
summary-schemas model 695
supply context 129
support 703
survey 592
synchronicity 730
synchronization 446, 666
synchronous 1090
  communication 525
  distance delivery 225
system 868
  -initiated WES 1102
systems 157

T

T1 460
T-1 access 908
TA or timing advance 636
TACACS (Terminal Access-Controller Access-Control System) 908
tag 115
tagged value 677
TAM (Technology-Acceptance Model) 341
tamper 212, 769
TDMA 973
TDOA or Time Difference of Arrival 636
teaching strategy 820
technologicalrisks 894
technology-mediated instruction 886
teleconferencing 225
telepresence 730
television 723
  commerce (t-commerce) 984
telework 950, 956
TELNET 868
Telnet 1037
TELRIC 544
temporal coordination mechanism 525
terminal defence 902
threats 403, 1000
time constraints 741
TOA or time of arrival 637
tokens 446, 666
topological relation 264
TPB (Theory of Planned Behaviour) 341
TRA (Theory of Reasoned Action) 341
trail 446
training 936
transactional software 246
transcoding 1007, 1077
translation 950
  -mediated communication (TMC) 950
transport layer 659
trap doors 956
treatment proxy 736
trojan horse 410, 956
troposphere 157
tunneling 956
TV object 518
twisted pairs 48
TWQ (teamwork quality) 537
typeface 979
typography 979

U

ubiquitous
  commerce (u-commerce) 984
  computing 607, 709, 1020, 1026
UBS (Universal Serial Bus) Port 806
UDDI (Universal Description, Discovery, and Integration) 88
UDF 381
UM (Unified Messaging) 88
UMA 1007
UMTS (Universal Mobile Telecommunication System) 294, 637, 973
unicast 30
  stream 736

URI (Uniform Resource Identification) 924
URL (Universal Resource Locator) 1077, 1082, 1095
usability 246, 537, 1012, 1082
USENET 1037
user 709, 761
  interaction 446, 666
  interface 246, 347
  -centered design 1013
  -initiated WES 1102
UTAUT (Unified Theory of Acceptance and Use of Technology) 341
UTF-8 101
UUCP 1037

V

VA 916
value chain 436
  -added services 686
values 241
vector graphic 979
verification 75
very-high-speed DSL (VDSL) 54
virtual
  class, virtual lecture, virtual seminar, virtual tutorial 1068
  community 1046, 1063
  knowledge space (Web site) 1054
  learning
    community 1063
    environments (VLE) 256
  private networks (VPNs) 956
  reality 374, 950
  school, virtual college, virtual university 1069
  team 241, 950
  universities 225
  /online community 841
virtuality 358
virus
  definition file (subscription service) 570
  signature 570
viruses 956
visibility 101
VLE (Virtual Learning environments) 537
voice
  commerce (v-commerce) 984
  over IP (VoIP) 328
  recognition 506
VRML 179, 862
vulnerabilities 403

W

W3C (World Wide Web Consortium) 924
WAAS 157
WAN 600
WAP (Wireless Application Protocol) 301, 644, 973, 1020, 1026
WAP Gap 621
watermark attack 212
WBT 741
WCDMA 973
WCDMA-FDD (Wideband Code-Division Multiple Access, Frequency-Division Duplex) 323
WDM 973
wearable computing devices 709
Web (World Wide Web, WWW) 1054
  cache 55
  conferencing 225
  impact factor 1095
  marketing 109
  page 1054
  search engine 207
  student 784
  -based
    systems 179
    teaching 979
webometrics (or internetometrics or cybermetrics) 1095
WEP 1113
wide area
  networks (WANs) 956
  radio network 651
Wifi 628
wired equivalent privacy (WEP) 621
wireless 659
  commerce (mobile or m-commerce) 984
  emergency service (WES) 129, 1103
  local community (WLC) 129
  transport security layer (WTSL) 644
WLAN (Wireless Local-Area Network) 367
WML 1020, 1026
WMLScript 1020
work for hire 790
workstation 761
World Wide Web (WWW) 423
worm 410
WPA 1113

WSDL (Web-Services Description Language) 88
WYSIWYG Visual Design Tools 179

X

X3D 703, 862
X-Media 578
XML (Extensible Markup Language) 115, 924, 1077
Xpointer 1077
XSL 1077
XSL style sheet 1032

Z

zoom 115
ZPD 95

Index

Symbols

(DoS) 748
(QoS) routing 22
3D manipulation 1031
4G networks 964
802.11i 1109
802.1x 1105

A

acceptable-use policy (AUP) 905
access 461
  control 849, 905
  networks 283
accountability 409
accounting Rates 501
acquisition 63
action units (AUs) 309
active learning 778
activity logging 904
ad hocs wireless network 605
ad hoc networks 646
adaptation 1001
  framework 1070
  techniques 1070
adaptive learning 974
adopt 1
advance mobile phone service (AMPS) 165
advanced distributed learning (ADL) 927
AdWord 267
affective computing 8, 311
AFX 856
agent (mobility) 18
Alan Kay 744
Alexandr Rodchenko 745
algorithms 22
alt 302
analog signal 180
animation 1057
anomaly detection 495
antivirus software 996
APON 331
application
  providers (ASPs) 31, 159
  software 995
  tier 757
Application-Layer Multicast (ALM) 875
archive 863
ARPANET 418
artificial intelligence 490
  bots and agents 408
ASCII 947
ASP 31
Association of Information Systems (AIS) 717
asymmetric digital subscriber line (ADSL) 42
asynchronous 882, 1057
  agent-to-agent communication 19
  digital-subscriber-line 418
  learning networks 417
  asynchronous transfer mode (ATM) 49, 359
ATM 830
  Forum 51
  protocol stack 51
AU coding of face images 310
auditing 995
auditory icons 344
augmented Reality 368
authentication 63, 215, 409, 849
  without identification 65
authorisation 409
authoritarian states 865
authorization 63, 849
automatic
  facial affect analyzers 10

Copyright © 2005, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

vocal affect analyzers 11
autonomous 15
Aviation Industry CBT 926
awareness and consent 65

B

B2B 278
B2C 278
B2E 278
back channel 428
back-propagation Network (BPN) 205
BBA 856
believable talking head (avatar) 308
Bernhard Riemann 743
BIFS 856
bimodal human-affect analyzer 12
binary image 763
biometric(s) 69
data 63
readers 997
systems 63
technologies 56
template 64
birds of a feather 1058
block data hiding method (BDHM) 763
Bluetooth 646
bottom-up development 937
BPON 331
brand 98
broadband 198, 283, 329, 418
access
solutions 76
technologies 78
integrated-services data networks (B-ISD 360
Internet connections 998
networks 22
browser 864
brute-force attacks 405
BSD license models 786
buffer overflow 998
bulletin board system (BBS) 937
business
games 532
model 557
process modelling 669
strategy 413
-to-Business 242

C

C2B 278
C2C 278
cable modem 418
captions 302
carrier-neutral collocation facilities 159
carrierless amplitude-phase (CAP) 43
censorship 864
certificate 1106
authority 850
Cézanne 743
chat rooms 1055
circuit-switched 318, 593, 646
class of service (CoS) 51
Claude Bragdon 745
client/server 756
cliques 1058
CMC 237
coaction fields 368
cognit 243
cognitive psychology 705
cohort 1058
collaboration 1056
collaborative
environments 368
learning 375
collocation 548
color 1080
communication 417
facilitators 89
model 946
systems 428
-based model 89
technologies 836
communities of practice 1055
of value 338
community 336
and Civic Networks 937
antenna systems 199
comparison 63
compressed edge-fragment sampling 751
computer
conferencing 139
mediated communication (CMC) 837
security 993
-aided translation 946
-based (CBT) 738
simulations 532

-mediated communication (CMC) 946
facilities 487
-mediated or computer-based communication 417
concept map 707
condominium fibre 161
conference calls 1055
confidentiality 408, 997
confidentiality, integrity, and availability (CIA) 396
connectedness 1042
constructivism 705
constructivist 130
consumer attitude 102
contact 462
content 513
adaptation 1070
integrity verification 213
management 110
provider/merchant (CP/M) 617
repurposing 110
-based access 36
-based multimedia retrieval 116
-based retrieval 711
-management facilities 487
-oriented model 89
context
-awareness 123
-sensitive human-affect analysis 12
continuous-time signal 180
control measures 890
controlled vocabulary 303
controls 371
converged services 593
convergence 200, 571
factor 575
index 576
measurement 574
convolution 182
cookies 477, 481
copyleft 787, 792
copyright 786
corporate conferencing 137
cost 144
and price rules 548
methodologies 144
-benefit analysis 501
counter propagation network (CPN) 205
course delivery 220
creation support 89
critical digital-mass index 575
critical mass 1
cross-discipline research 756
CSCL 375
CSCW 1023
cued keyword personal questions 63
cultural
and personality differences 244
communication expectations 461
factors 461
transformation 91
values 234
cursor 743
customer-owned dark-fibre networks 160
cutoff frequency 186
cyberspace 417

D

DAB (digital audio broadcasting) 199
dark-fibre infrastructure 159
data 220
broadcasting 199
carousels 199
collection 480
integrity 993
mining 302, 475, 696
overload problem 696
piping 199
recovery 995
dtreaming 199
dedicated services 431
deep tissue illumination 64
delegation of authority 849
delivering content 1001
denial-of-service (DoS) attacks 406, 748, 905
denial-of-service 748, 998
derived work 786
Descartes' 743
detection methodologies 494
developing countries 447
dialect 464
dictionary attacks 405
difference equation 182
differentiated services (DiffServ) 869
diffusion 1
diffusive systems 428
digital
certificates 850
filter 181
item adaptation (DIA) 1005

mock-up 1028
Pearl Harbour 408
representation 382
signal processing (DSP) 180
signals 180
signatures 997
subscriber-loop (DSL) 547
television 197, 678
video-broadcasting 197
watermarking 213, 382
direct form 182
directed instruction 135
discrete multitone (DMT) 43
discussion forums 1055
distance
education 219, 447, 454
education systems 454
teaching 738
distanced leadership 226
distributed DoS 749
distribution 807
DNA fingerprint 64
Douglas Engelbart 744
DVB 678
dynamic multimedia proxy scheme 731

E

e-banking 279
e-businesses 158
e-commerce 96, 623, 980
e-economy 158
e-investments 279
e-learning 49, 271, 417, 925, 978
e-marketplace 887
e-shopping 279
E2E 278
EAP 1105
earcons 344
ease-of-use 987
economic risks 889
EDGE 291
edge-sampling method 751
education 738
EFM Fiber 330
Einstein 743
elaboration likelihood model (ELM) 265
electronic
banking 279
businesses 158
commerce 96, 102, 623, 980
customer relationship management (eCRM) 667
economy 158
investments 279
learning 49, 271, 417, 738, 925, 978
mail 418, 1055
marketplace 887
media 353
Program Guide (EPG) 431
shopping
embedding distortion 213
emergency call service (ECS) 1096
emergent behavior 16
emotional facial expressions 310
employees or humans 997
encryption 65, 997
enhanced data for GSM evolution (EDGE) 360
enhanced prioritized ptri net 660
enhanced-definition television (EDTV) 198
enrollment 63
entertainment 102
Ethernet 283
PON 332
Euclidean geometry 743
evolutionary 257
ex ante predictors of student performance 778
exchangi 488
exposures 910
expressions of emotions 9
eXtensible Markup Language (XML) 173
extraction 63

F

f security managem 494
face
recognition 57
-based command issuing 308
-to-face 1055
facial
recognition 72
scanner 64
facilitator 1058
factory tool 930
fair information practices (FIP) 476
false acceptance 66
FBA 856
fiber
optics 220
-to-the-home (FTTH) 329

fibre
  to the building (FTTB) 159
  to the curb (FTTC) 159
  to the home (FTTH) 159
filesharing 807
fingerprint
  biometric 57
  scanner 64
  scanning 71
firewalls 407, 996
first generation 165
  cellular systems 290
five-component model 756
flexibility 200
fluency 463
fluropolymers 830
forking source code 794
forward error correction 44
four-dimensional geometry 742
Fourier transform 183
fourth dimension 744
fragile watermarking 215
free
  software 785
  speech 863
freedom of information 865
FreeNets movement 937
frequency
  division multiple access (FDMA) 165
  response 183
  sampling 186
  scalability 1002
  -division multiplexing (FDM 43
FTTx schemes 159
fundamental 144

G

G2C 278
gain function 184
Galileo 152, 348
gatekeepers 866
general packet radio service (GPRS) 360, 647
general public license (GPL) 555
genetic load balancing routing (GLBR) 22
Gigabit PON 332
global system for mobile communications 647
global virtual teams (GVTs) 519
globalization 947
globalize 96
GLONASS 151, 348
GNSS 348
GNU General Public License models 786
Google 265
GPRS 291
GPS 151, 348
granularity 799
graphical
  interfaces 864
  user interfaces 742
grayscale images 204
GSM 143, 290, 647
  -based (global system mobile) 359

H

hackers 407, 997
hand geometry 64
handover 653
hardware 995
health-care industry 396
Hermann Minkowski 743
heterogeneous wireless networks 652
hierarchical caching architecture 732
high
  altitude stratospheric platform (HASP) 359
  -definition TV (HDTV) 198
  -level features 117
  -speed circuit-switched data (HSCSD) 647
higher-level learning 705
home
  networking 76
  networks 78
host-multicast 875
HSCSD 291
HSDPA 320
HTML 302
HTTP 864
human
  intelligence (HI) 1064
  -computer interaction (HCI) 8, 532, 1008
hybrid fibre coaxial (HFC) 159
HyperClass 1066
hypermedia systems 173
HyperReality 368, 1064
hypersolids 745
hyperspace of the parameters 876
hypertext 742, 863, 1057
HyperWorld (HW) 368

I

i-mode 296
idea generation 1047
identification 63
identity theft 998
idiomatic expressions 464
IDS Deployment 494
IDSs 996
IEEE (Institute of Electrical and Electronics Engineering) 1104
IEEE Learning Technology Standards Committee (LTSC) 926
image retrieval 257
implementation 467
  risks 888
incentive regulation 538
individual
  critical mass 1
  passwords 63
  styles 1047
infinite impulse response 181
info pyramid 1001
information 993
  and communication technologies (ICTs) 353, 467, 1055
  containers 89
  hiding 382
  management 507
  retrieval (IR) technologies 710
  security 63, 404, 396, 842, 895
    management 390, 895, 993
    systems 31, 756
  visualisation 1021
  -objects 743
  -security-management system (ISMS) 396
informativeness 102
informed embedding 215
infrastructure 758
inhabitants 369
inite impulse response 181
innovation diffusion theory 1086
innovative nature 200
input device 706
insider attacks 407
installed base 1
instant messaging 418
instructivist 131
Integrated Services (IntServ) 869
integration issues 888, 889
integrity 63, 408
intellectual
  property 785, 792
  wealth 864
intelligent 22
  agent 706
interaction 417
  channel 198
  -oriented individuals 1058
interactive 865
  advertising 431
  games 431
  multimedia technologies 447
  systems 428
  Web applications 84
interactivity 103, 200, 576, 724
intercultural communication 236
interdependencies 889
interface 706
interfacing 200
International Organization for Standardization 1014
international
  outsourcing 795
  virtual office (IVO) 461
internationalization 947
internationalizing 97
Internet 417, 585, 863, 1034, 1055, 1091
  adoption 467
  privacy 480
  service provider (ISP) 159, 418, 550
  -based systems 910
interoperability 199, 489
interorganizational 526
intrusion
  attacks 405
  detection 494
    system (IDS) 494, 749, 996
  prevention 749
  tracing 749
investment strategy 500
ionosphere 153
IP
  multicasts 50
  traceback 749
  -multicast 875
    routing protocol 875
IPSec 1111
IPsec 84

IPsec authentication 753
IPv4 653
IPv6 653
iris scanning 57
irrevocability 67
IS management 507
ISO 1008
ISO17799 396
IT 508
  alignment 411
  capabilities 508
  expertise 508
  strategy 411
  success 508
iTV 512
Ivan Sutherland 744

K

Kaiser Permanente 419
keystroke pattern 64
knowledge 353, 467
  domain 371
  management 526, 1047
  -construction model 89

L

LAAS 153
LAN 1104
last mile 76, 543
laws 371
leadership 237, 509
  competencies 519
learning 532
  environment 1060
  object metadata 247
  objects 249, 486, 882
  process 738
  styles 706, 883
  -object metadata (LOM) 487
Lesser GNU Public License 787
life cycle 1042
linking pages 742
lip reading 308
local
  access networks 547
  competition 538
  interactivity 429
  loop unbundling 538
  loop unbundling (LLU) 547
  network 538
  telecommunications 538
  -area networks (LANs) 160
localization 946
localize 97
location
  tracking 622
  transparency and dependency 623
  using positioning technologies 630
  -based applications 629
  -based services 629
lossy compression 215
low-level features 117
lowpass filter 186
Ludwig Wittgenstein 745

M

m-commerce 279, 342, 623
m-learning 704, 1024
MAC 602
machine
  learning 310
  translation 945
  vision 310
  -context sensing 12
magnitude response 183
make or buy 31
malicious software 405
malware 406
managed learning environments (MLEs) 248, 487
managing information security 842
market-led nature 200
maturity 509
means of communication 371
mechanical simulation 424
media search engines 302
medical application 257
Memex 437
message-based service 579
messaging 752
metadata 306
Microbrowser 110
microwave signals 220
MIMO 320
MIS Classrooms 717
misuse detection 495
Mobey Forum 619
mobile 1106
mobile ad hoc network (MANET) 601

mobile
agent 608
banking 295
business 624
commerce 123, 295, 615, 638, 982
security 615
communications 290, 315, 629
computing 622, 1016
connectivity 623
device 622, 638, 704, 1036
electronic transactions (MeT) 619
HCI 1021
multimedia computing 638
networks 143, 964
payment 616, 623
forum 619
security 623
systems 964
transactions 623
users 622
value-added service 583
MOBILearn 704
mobility 86, 652
modality 343
mode of participation 1041
model 756
modes 343
modification 888
monitoring 995
motivations 1041
mouse 743
Mozilla Public License (MPL) models 786
MP3 696
MPEG 696, 855
MPEG-2- (Moving Pictures Experts Group-2) 197
MPEG-21 1005
MPEG-4 770
MPQoS 770
MSOs 324
multi-
agent systems (MAS) 15
dimensional indexing 119
modality 713
Protocol Label Switching (MPLS) 869
multicast
strategy 18
transport protocols 875
multimedia 130, 660, 731, 742
communication 315
content representation 687
technologies 687
copyright protection 213
database Systems 273
databases 271
document(s) 710, 1001
home platform (MHP) 201
information retrieval 710
instructional materials 717
interactivity 724
materials 704
networking 965
security management 214
services 678
technologies 123, 737
multimodal 346
multipath fading 316
multiple system operators 324
multiple-channel standard-definition TV (SDTV) 198
multiplexing 645
multiprotocol
encapsulation 199
label switching (MPLS) 50
multiuser
domains (MUDs) 419
-domain object oriented (MOO) 419
municipal fibre network 161
museum 739

N

narrowband 329
national regulatory authorities (NRAs) 548
natural
human-machine interaction 311
language 945
naturalness 961
navigate 743
navigation 1080
negotiation 1056
Net Present Value (NPV) 501
netizens 865
network 756, 910
effects 1
of networks 418
traffic 742
neural network 204
news organizations 865

Nicholas Lobatchevsk 743
Node appending 751
non-recursive 182
non-repudiation 63
non-verbal communicative signals 9
Nonrepudiation 997
nonverbal communication 948
numeric addressing 866

O

objective PQoS evaluation 770
OBS 799
OELE 136
One-Way Interactivity 429
onfidentiality 63
online 425
  communities 336, 836, 937
  discussion board 778
  education 977
  forums 417
  media 461
ONT 326
OoBE 985
open
  knowledge initiative 247
  source 555
    definition 786
    initiative 786, 791
    license 786
    software 555, 785, 791
openness 199
operating system 791
opt-in/opt-out 477
optical 799
  burst switching 799
  Internet 799
  line terminal (OLT) 330
  network terminal (ONT) 326, 330
  -access networks (OANs) 158
optimization models 83
Organization for Economic Co-operation and Develop 477
organizational memory 1058
OSI Model 756
out-of-the-box experience 985
outsourcing 464

P

P2P 278
packet
  marking 751
  sniffing 865
  switched networks 593
  switching 593
  -switched 318
palm print 58
passband 186
  edge frequency 186
  ripple 186
passive
  optical networks 283, 331
password 994
  attacks 405
  cracking 405
  sniffing 888
patches 997
patterns 866
pay per view 431
payback rules 501
payload 406
payment
  authentication 297
  service provider (PSP) 617
PDAs 110, 1024, 1036
peer review method 375
peer-to-peer (P2P)
  file-sharing systems 807
  network 601
perception 243
peripheral cues 267
personal computers 418
personnel controls 994
Petri net 437, 660
phase 183
phishing 405
phonemes 959
physical
  reality (PR) 1064
  security 994
physiology 243, 244
Picasso 743
pico-network 646
picture-archiving 396
  and communication system 821
Pierre Fermat 743

PIN identification 63
pixel 762
PKI 1107
plain, old telephone system (POTS) 42
planning 509
plastic optical fiber (POF) 829
platform penetration 573
plug-and play 985
poles 184
polycarbonate core material (PC) 830
polymethylmetacrylate core material (PMMA) 830
polystyrene core material (PS) 830
post 866
  -Euclidean 743
power
  control 319
  distance 463
  law 867
PQoS 770
present conferencing 139
presentation 757
principal component analysis (PCA) 118
prioritized Petri Net 660
privacy 65, 475, 480
  seal 477
privilege management infrastructure (PMI) 849
proactive tracing 751
probabilistic neural network (PNN) 205
probing or scanning 405
profile management 625
profitability 144
program-related services 431
programmer 792
prometeus 927
proprietary software 791
prosody 960
protocol 802
  parameters 876
prototyping 424
proxy cache 731
public
  key infrastructure (PKI) 297
  -safety answer points (PSAPs) 1096

Q

QBE (query by example) 36
quadrature amplitude modulation (QAM) 43
quality assessment 1004
quality of service (QoS) 869
quality scalability 1002
quality-of-service (QoS) 49, 652
quantization index modulation (QIM) watermarking 215
query 732
query-by-example (QBE) 12, 119

R

radiology 821
reactive
  learn 131
  tracing 753
recursive
  part 182
  running–sum filter 184
reflective practice 133
region of convergence 184
regression analysis 501
regulation 297
relational risks 889
relevance feedback 119
reliability 875
Remez method 186
remote procedure call (RPC) 608
remote users 956
representation technologies 687
representative frames 112
repudiation 888
research methods 585
resource reservation protocol (RSVP) 869
retina and iris scanning 70
retinal scanning 58
retinal/iris scanner 64
reversible watermarking 216
RFID 1022
rich media 84
risk
  analysis 994
  management 787
  -control framework 892
risks 66
robust watermarking 214
routing protocol 603

S

salient objects 257
same-time, same-place setting 417
Sarbanes-Oxley Act 994

satellite navigation 348
scalability 889
scalable coding 1001
scatter-networks 646
SCO-IBM lawsuit 787
search
  engine 97
  technologies 742
second-generation (2G) cellular systems 291
secure-sockets layer (SSL) 84
security 390, 895, 910, 952, 1104
  event management 999
  laboratory 903
  log 997
  loopholes 85
  management 390, 397
  standard 1111
  threats 390
self-modifiable color Petri Net 437
self-oriented individuals 1058
semantic
  gap 120
  Web 86, 917
semantic web 249
semantics 1003, 1031
semi-fragile watermarking 215
sensor network 602
service
  continuity 995
  level agreement (SLA) 872
services 512
set top box or decoder 430
SGML 864
shareable content object reference model (SCORM) 247, 487
shared workspace 1057
signal 180
  processing 180
similarity
  matching 36
  measurement 118
simulation 1057
singularities 184
site specification 97
small firm 413, 467, 509
smurf attack 749
SNHC 856
sniffer programmes 405
SNITCH protocol 752
social
  capital 1058
  communication 91
  constructivism 1055
  engineering 407, 99
  relationships 419
  -constructivism paradigm 89
socially constructed reality 353
sociological 585
Socratic dialogue 778
software 996
  agents 15, 911
  patent 787
soundscape 345
source code 791
spam 481
spatial 257
  scalability 1002
speech synthesis 957
splitter 550
spoofing 66, 405, 888
spread spectrum (SS) 316
  watermarking 214
  techniques 646
spyware 406
stability 182
status 462
steganography 382, 762
stopband 186
  attenuation 186
  edge frequency 186
  ripple 186
storage 63
streaming 110
subscriber identity module (SIM) 616
supply chain management 888
switched digital video (SDV) 159
synchronization 660
synchronous 1057
  agent-to-agent communication 19
synthesizers 958
system 910
  function 184
  software 995
  -initiated WES 1096

T

tactile-kinetic motion 742
tags 111

talon 882
task-oriented 1058
TCP SYN flooding 749
teamwork quality construct (TWQ) 532
technological risks 888
technology-integrated learning systems 455
Ted Nelson 742
telcos 324
telecommunication 144
  companies 324
telework 951
TELRIC 540
temporal
  relations 258
  scalability 1002
text preprocessing 959
text-to-speech 957
The ARIADNE Foundation 927
the future of Web-based learning 1088
the history of Web-based learning 1084
Theo Van Doesburg 744
theory
  of consumption values 338
  of planned behavior 1086
therapeutic 821
three-dimensional 743
Tim Berners-Lee 418
TKIP 1105
token-based arrangements 63
TPB 1086
training 925, 995
transaction costs 888
transcoding 1001
transition 186
translation 945
  memory 948
transmissive power 867
transmoding 1003
Trojan horses 406
troposphere 153
trust 236, 1056
trusted third party (TTP) 617
TV banking 432
twisted pairs of copper wires 42
two-way interactivity 429
type justified 974

U

u-commerce 279, 980
ubiquitous 605
  commerce 980
  computing 704, 1025
UDP-flood attack 749
UMTS 315
undergraduate IS classrooms 718
unified modeling language (UML) 668
uniform resource identification 919
universal 863
  mobile telecommunication service (UMTS) 647
  mobile telecommunications system (UMTS) 359
  multimedia access (UMA) 1001
usability 1014, 1008, 1078
usage environment 1002, 1005
user
  interface 342
  satisfaction 1001
  -initiated WES 1096
UTF-8 100

V

video 220
  conferences 1057
  conferencing 138
  data similarity 36
  on demand 327
  surveillance 8
  -indexing and -retrieval 36
  -on-demand (VOD) 159
  -retrieval system 36
virtual
  class 1065
  clothing 424
  community(ies) 337, 417, 585, 836, 937, 1033, 1040, 1055
  learning community 1055
  learning environments (VLEs) 248, 487
  organizations 226
  private networks (VPNs) 951
  reality (VR) 367, 948, 1064
  teams 233, 417
  worlds 1036
virtuality 353
visibility 96
visualization 1057
  of geometry 743

VOD 327
voice 220
voice over IP 324
voice tecognition 58, 72
voiceprint 64
VoIP 324
voluntary cooperation 297
VRML 856
vulnerabilities 910, 1105

W

WAAS 153
walk through 743
WAP (Wireless Application Protocol) 296, 1021
  gap 617
watermarking 204, 382
WCDMA 315
wearable communities 1036
Web
  access 500
  application 426
  bugs 481
  collaboration 83
  community 937
  conferencing 86, 138
  logs 865
  site 1078
  student 778
  usability 1010, 1078
  -based (WBT) 738
    distance learning 882
    hypermedia systems 173
    learning 1084
    teaching/learning 974
webometrics 1091
  procedures 1092
weighted-least-squares 186
WEP 1111
Wi-Fi 1104
wide-area 647
window methods 186
  ired Equivalent Privacy (WEP) 617
wireless 1106
  access protocol 624
  application protocol (WAP) 648
  commerce 980
  communications 623
  device 295
  emergency call service 1096
  local area networks 646, 998
  middleware 625
  mobile devices 708
  network(s) 123, 622
  technology 165, 500
  terminals 970
WLANs 165, 646, 998
WML 1024
World Wide Web 418, 697, 742, 917, 1055
worm/virus 405
WWW 917, 1091

X

X3D 856, 1030
Xerox's Palo Alto Center 744
XML 112, 1030

Z

z- transform 183
zeros 184
zone of proximal development 90, 1061
zooming 111